



US 20110016135A1

(19) **United States**

(12) **Patent Application Publication**  
**Teerlink**

(10) **Pub. No.: US 2011/0016135 A1**

(43) **Pub. Date: Jan. 20, 2011**

(54) **DIGITAL SPECTRUM OF FILE BASED ON CONTENTS**

(76) Inventor: **Craig N. Teerlink**, Cedar Hills, UT (US)

Correspondence Address:  
**KING & SCHICKLI, PLLC**  
**247 NORTH BROADWAY**  
**LEXINGTON, KY 40507 (US)**

(21) Appl. No.: **12/616,306**

(22) Filed: **Nov. 11, 2009**

**Related U.S. Application Data**

(60) Provisional application No. 61/236,571, filed on Aug. 25, 2009, provisional application No. 61/271,079, filed on Jul. 16, 2009.

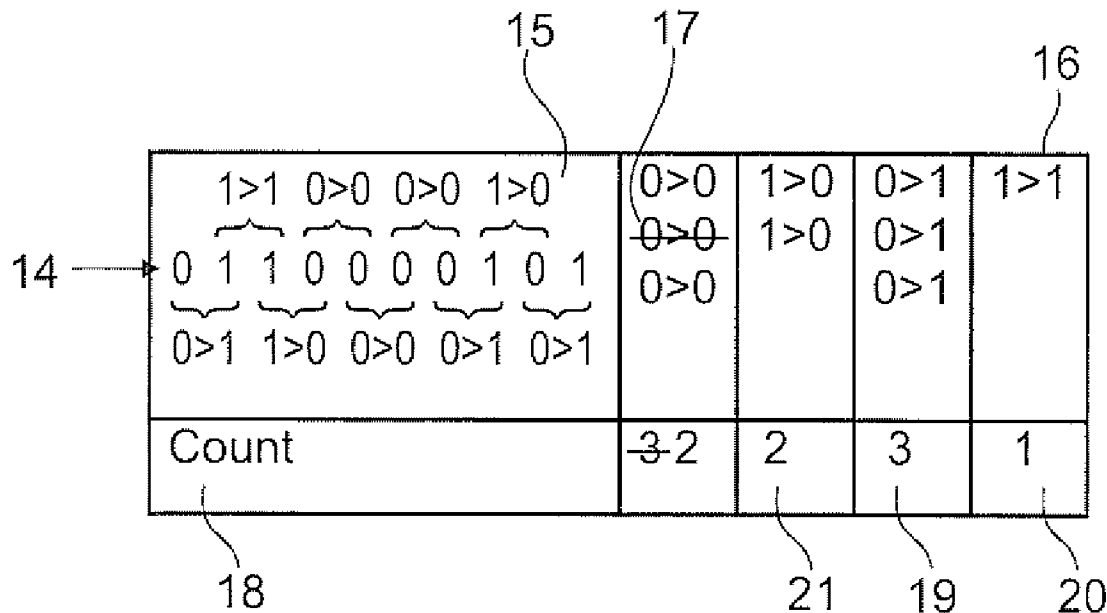
**Publication Classification**

(51) **Int. Cl.**  
**G06F 17/30** (2006.01)

(52) **U.S. Cl. ... 707/749; 707/748; 707/736; 707/E17.033**

(57) **ABSTRACT**

A digital spectrum defines and communicates a file's informational characteristics. A file's informational position may be represented as a vector in an N-dimensional space, where each dimension is defined by a symbol. A position along the axis of any given dimension is described by the frequency of occurrence of that symbol. Relative to the origin of the space, the file's position can be computed. Comparing positions reveals similarity, or not, of the files. Another method uses the digital spectrum to define a line graph. Each symbol and its frequency define points on the line. A distance function between two spectra line graphs is computed. Comparing values from the distance functions reveals similarity, or not, of the files. Also, total numbers of bits in the files are extracted by knowing the lengths of the original bits corresponding to every symbol. A symbol bit length spectrum is also defined.



10

Term	Definition
Alphabet	The set of all possible symbols currently in use.
Atomic symbols	The symbols 0 and 1, which are based on the raw 0 and 1 bit values. All subsequently defined symbols represent tuples that are based on symbols 0 and 1.
Character	A symbol that appears in the data stream.
Compressed file size	The number of bits that are required to store the encoded data stream, the dictionary, and the symbol-encoding information.
Data stream	A sequential stream of characters. The terms data stream and text are synonymous in this document.
Dictionary	A collection of information regarding all symbols (the alphabet).
Encoded data stream	A data stream of Huffman encoded characters.
Pass	The performance of one iteration of the compression procedure applied to the current data stream.
Symbol	A unit of information. The information that is represented by a symbol can be from 1 to N binary bits in length.
Symbol-encoding information	Symbols in the alphabet are digitally encoded to reduce the amount of space required to store or transmit them electronically. The encoding information is stored and used to decompress the data later.  A well understood method of minimizing the space required to store a series of characters is the use of minimum weighted path length trees, as given by David Huffman (D. E. Knuth, <i>The Art of Computer Programming</i> , 1973, vol. 1, p. 402).
Text	A sequential stream of characters. The terms data stream and text are synonymous in this document.
Tuple	Two adjoining characters in the data stream or text. The order of the appearance of characters in the tuple is designated as "first" and "last". The notation for tuples is "first>last" to show the order of appearance in the pair of characters and to avoid confusion of the tuples with real numbers. For example, a tuple of symbol 1 followed by symbol 0 is written as 1>0.  In each pass through the data stream, the most highly occurring tuple is determined. A new symbol is created to represent the tuple in the data stream. The symbol stands for and replaces all occurrences of the tuple in the data stream.

FIG. 1

12

Tuple Array	First	
Last	0	1
0	0>0	1>0
1	0>1	1>1

FIG. 2

15      17      16

14 →	1>1 0>0 0>0 1>0	0>0	1>0	0>1	1>1
	0 1 1 0 0 0 0 1 0 1	0>0	1>0	0>1	
	0>1 1>0 0>0 0>1 0>1	0>0		0>1	
Count		3	2	3	1

18      21      19      20

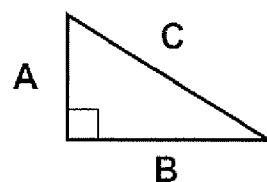
FIG. 3

22

Tuple Count	First	
Last	0	1
0	2	2
1	3	1

19

FIG. 4



Pythagoreean Theorem

$$A^2 + B^2 = C^2$$

FIG. 5

23

Tuple	A <sup>2</sup>	B <sup>2</sup>	C <sup>2</sup>	Hypotenuse
1>5	1	25	26	5.1
3>7	9	49	58	7.6
4>4	16	16	32	5.7

FIG. 6

24

Tuple Array	First							
Last	0	1	2	3	4	5	6	7
0	0>0	1>0	2>0	3>0	4>0	5>0	6>0	7>0
1	0>1	1>1	2>1	3>1	4>1	5>1	6>1	7>1
2	0>2	1>2	2>2	3>2	4>2	5>2	6>2	7>2
3	0>3	1>3	2>3	3>3	4>3	5>3	6>3	7>3
4	0>4	1>4	2>4	3>4	4>4	5>4	6>4	7>4
5	0>5	1>5	2>5	3>5	4>5	5>5	6>5	7>5
6	0>6	1>6	2>6	3>6	4>6	5>6	6>6	7>6
7	0>7	1>7	2>7	3>7	4>7	5>7	6>7	7>7

25

FIG. 7

FIG. 10

40

Tuple Count	First	
	0	1
Last	0	1
0	95	96
1	96	48

FIG. 11

41

Symbol	Count
0	240
1	144
Total	384

FIG. 12

26<sup>1</sup>

Alphabet	Definition	
	First	Last
0	0	-
1	1	-
2	1	0

FIG. 13

FIG. 14

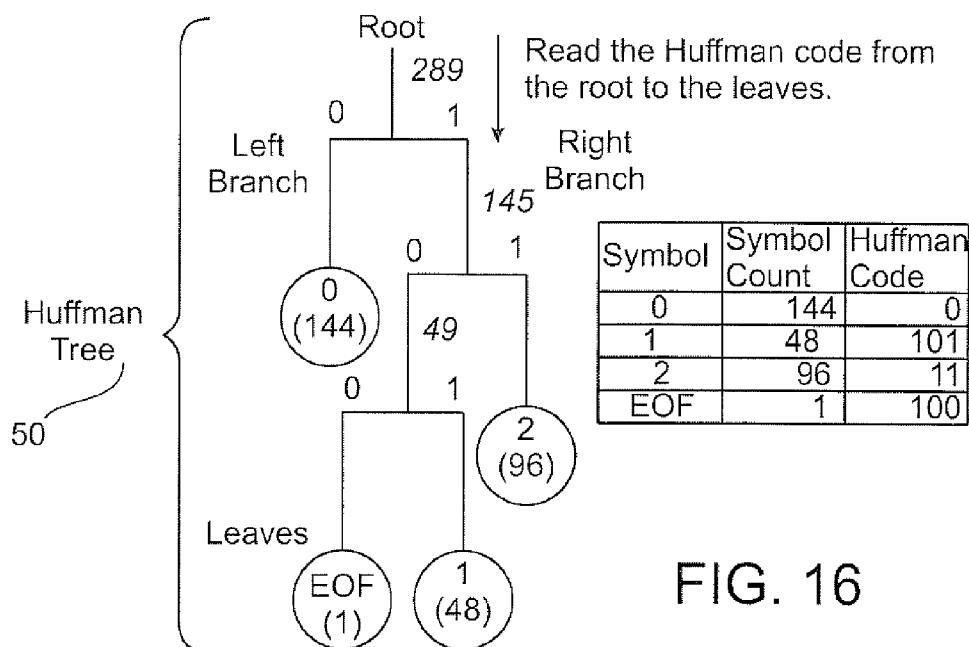
30<sup>1</sup>

0 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1  
 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0  
 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0  
 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1  
 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0  
 0 0 2 1 2 0 0 2 0 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 2  
 0 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1  
 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 0 2 1 2 0 0 2 0 1 2 0 0 0 2 1 2 0  
 0 2 0 1 2 0 0 2 0 1 2 0 0 2 0 1 2 0 0 2 0 1 2 0 0 2 0 1 2 0 0 2

FIG. 15

Symbol	Count
0	144
1	48
2	96
Total	288

41<sup>1</sup>



Symbol	Huffman Code	Bit Length	Symbol Count	Total Bits
0	0	1	144	144
1	101	3	48	144
2	11	2	96	192
EOF	100	3	1	3
Total Bits for Data				483

FIG. 17

Compression Overhead	Current
File Information	8
Dictionary Length	2
Tree Length	15
Total Overhead	25

FIG. 18

Compressed File Size	Original Bit Count	Current Bit Count
Data Length	384	483
Overhead	0	25
Total Bits Needed	384	508
Compression Ratio	132%	

FIG. 19



Tuple Array	First		
Last	0	1	2
0	0>0	1>0	2>0
1	0>1	1>1	2>1
2	0>2	1>2	2>2

35<sup>1</sup>

FIG. 20

Tuple Count	First		
Last	0	1	2
0	48	0	56
1	9	0	39
2	48	48	0

40<sup>1</sup>

FIG. 21

Alphabet	Definition	
	First	Last
0	0	-
1	1	-
2	1	0
3	2	0

26<sup>11</sup>

FIG. 22

FIG. 25

57<sup>1</sup>

Compression Overhead	Current
File Information	8
Dictionary Length	6
Tree Length	24
Total Overhead	38

FIG. 26

58<sup>1</sup>

Compressed File Size	Original	Current
Data Length	384	507
Overhead	0	38
Total Bits Needed	384	545
Compression Ratio	141%	

FIG. 27

35<sup>11</sup>

Tuple Array	First			
Last	0	1	2	3
0	0>0	1>0	2>0	3>0
1	0>1	1>1	2>1	3>1
2	0>2	1>2	2>2	3>2
3	0>3	1>3	2>3	3>3

FIG. 28

40<sup>11</sup>

Tuple Count	First			
Last	0	1	2	3
0	39	0	0	48
1	1	0	39	8
2	40	0	0	0
3	8	48	0	0

FIG. 29

26<sup>111</sup>

Alphabet	Definition	
	First	Last
0	0	-
1	1	-
2	1	0
3	2	0
4	1	3

FIG. 30

30-4  
↓

```

0 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0
2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0
2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0
2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 3 4 0 0 2 4 0 0 2 4 0 0 2 4 0 3 4
0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4 0 0 2 4
0 3 4 0 0 2 4 0 3 4 0 3 4 0 3 4 0 3 4 0 3 4 0 2

```

FIG. 31

41<sup>111</sup>  
↙

Symbol	Symbol Count	Huffman Code
0	88	0
1	0	-
2	40	111
3	8	1101
4	48	10
EOF	1	1100

FIG. 32

52<sup>11</sup>  
↙

Symbol	Huffman Code	Bit Length	Symbol Count	Total Bits
0	0	1	88	88
2	111	3	40	120
3	1101	4	8	32
4	10	2	48	96
EOF	1100	4	1	4
Total Bits for Data				340

FIG. 33

57<sup>11</sup>

Compression Overhead	Current
File Information	8
Dictionary Length	10
Tree Length	24
Total Overhead	42

FIG. 34

58<sup>11</sup>

Compressed File Size	Original	Current
Data Length	384	340
Overhead	0	42
Total Bits Needed	384	382
Compression Ratio		99%

FIG. 35

35<sup>111</sup>

Tuple Array	First				
Last	0	1	2	3	4
0	0>0	1>0	2>0	3>0	4>0
1	0>1	1>1	2>1	3>1	4>1
2	0>2	1>2	2>2	3>2	4>2
3	0>3	1>3	2>3	3>3	4>3
4	0>4	1>4	2>4	3>4	4>4

FIG. 36

40<sup>111</sup>

Tuple Count	First				
Last	0	1	2	3	4
0	39	0	0	0	48
1	0	0	0	0	0
2	40	0	0	0	0
3	8	0	0	0	0
4	1	0	39	8	0

FIG. 37

26-4

Alphabet	Definition	
	First	Last
0	0	-
1	1	-
2	1	0
3	2	0
4	1	3
5	4	0

FIG. 38

30-5  
↙

```

0 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5
0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2
5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 3 5 0 2 5 0 2 5 3 5
0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 0 2 5 3 5 3 5
5 3 5 3 5 3 5 2
  
```

FIG. 39

41-4  
↙

Symbol	Symbol Count	Huffman Code
0	40	10
1	0	-
2	40	01
3	8	001
4	0	-
5	48	11
EOF	1	000

FIG. 40

52<sup>111</sup>  
↙

Symbol	Huffman Code	Bit Length	Symbol Count	Total Bits
0	10	2	40	80
2	01	2	40	80
3	001	3	8	24
5	11	2	48	96
EOF	000	3	1	3
Total Bits for Data				283

FIG. 41



57<sup>111</sup>

Compression Overhead	Current
File Information	8
Dictionary Length	16
Tree Length	24
Total Overhead	48

FIG. 42

58<sup>111</sup>

Compressed File Size	Original	Current
Data Length	384	283
Overhead	0	48
Total Bits Needed	384	331
Compression Ratio		86%

FIG. 43

35-4

Tuple Array	First					
Last	0	1	2	3	4	5
0	0>0	1>0	2>0	3>0	4>0	5>0
1	0>1	1>1	2>1	3>1	4>1	5>1
2	0>2	1>2	2>2	3>2	4>2	5>2
3	0>3	1>3	2>3	3>3	4>3	5>3
4	0>4	1>4	2>4	3>4	4>4	5>4
5	0>5	1>5	2>5	3>5	4>5	5>5

FIG. 44

Tuple Count	First					
Last	0	1	2	3	4	5
0	0	0	0	0	0	39
1	0	0	0	0	0	0
2	39	0	0	0	0	1
3	0	0	0	0	0	8
4	0	0	0	0	0	0
5	1	0	39	8	0	0

25<sup>11</sup> FIG. 45

Alphabet	Definition	
	First	Last
0	0	-
1	1	-
2	1	0
3	2	0
4	1	3
5	4	0
6	2	5

FIG. 46

0 5 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6  
0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6 0 6 3 5 0 6 0 6 0 6  
3 5 0 6 0 6 0 6 0 6 0 6 0 6 0 6 3 5 0 6 3 5 3 5 3 5 3 5  
2

FIG. 47

41-5

Symbol	Symbol Count	Huffman Code
0	40	0
1	0	-
2	1	10100
3	8	1011
4	0	-
5	9	100
6	39	11
EOF	1	10101

FIG. 48

52-4

Symbol	Huffman Code	Bit Length	Symbol Count	Total Bits
0	0	1	40	40
2	10100	5	1	5
3	1011	4	8	32
5	100	3	9	27
6	11	2	39	78
EOF	10101	5	1	5
Total Bits for Data				187

FIG. 49

57-4

Compression Overhead	Current
File Information	8
Dictionary Length	22
Tree Length	29
Total Overhead	59

FIG. 50

58-4

Compressed File Size	Original	Current
Data Length	384	187
Overhead	0	59
Total Bits Needed	384	246
Compression Ratio	64%	

FIG. 51

35-5

Tuple Array	First						
Last	0	1	2	3	4	5	6
0	0>0	1>0	2>0	3>0	4>0	5>0	6>0
1	0>1	1>1	2>1	3>1	4>1	5>1	6>1
2	0>2	1>2	2>2	3>2	4>2	5>2	6>2
3	0>3	1>3	2>3	3>3	4>3	5>3	6>3
4	0>4	1>4	2>4	3>4	4>4	5>4	6>4
5	0>5	1>5	2>5	3>5	4>5	5>5	6>5
6	0>6	1>6	2>6	3>6	4>6	5>6	6>6

FIG. 52

40-5

Tuple Count	First						
Last	0	1	2	3	4	5	6
0	0	0	0	0	0	4	35
1	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0
3	0	0	0	0	0	4	4
4	0	0	0	0	0	0	0
5	1	0	0	8	0	0	0
6	39	0	0	0	0	0	0

FIG. 53

26-6

Alphabet	Definition	
	First	Last
0	0	-
1	1	-
2	1	0
3	2	0
4	1	3
5	4	0
6	2	5
7	0	6

FIG. 54

[illegible]

FIG. 55

Symbol	Symbol Count	Huffman Code
0	1	0100
1	0	-
2	1	01011
3	8	011
4	0	-
5	9	00
6	0	-
7	39	1
EOF	1	01010

FIG. 56

Symbol	Huffman Code	Bit Length	Symbol Count	Total Bits
0	0100	4	1	4
2	01011	5	1	5
3	011	3	8	24
5	00	2	9	18
7	1	1	39	39
EOF	01010	5	1	5
Total Bits for Data				95

FIG. 58

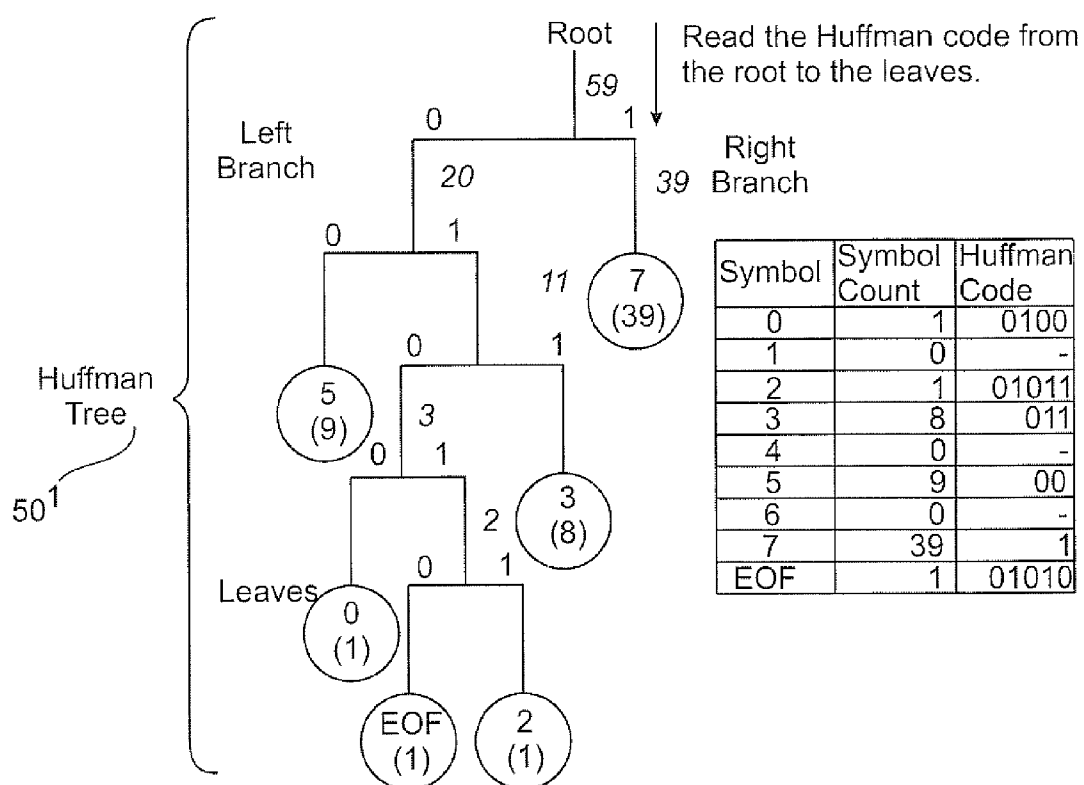


FIG. 57

57-5

Compression Overhead	Current
File Information	8
Dictionary Length	28
Tree Length	35
Total Overhead	71

FIG. 59

58-5

Compressed File Size	Original	Current
Data Length	384	95
Overhead	0	71
Total Bits Needed	384	166
Compression Ratio	43%	

FIG. 60

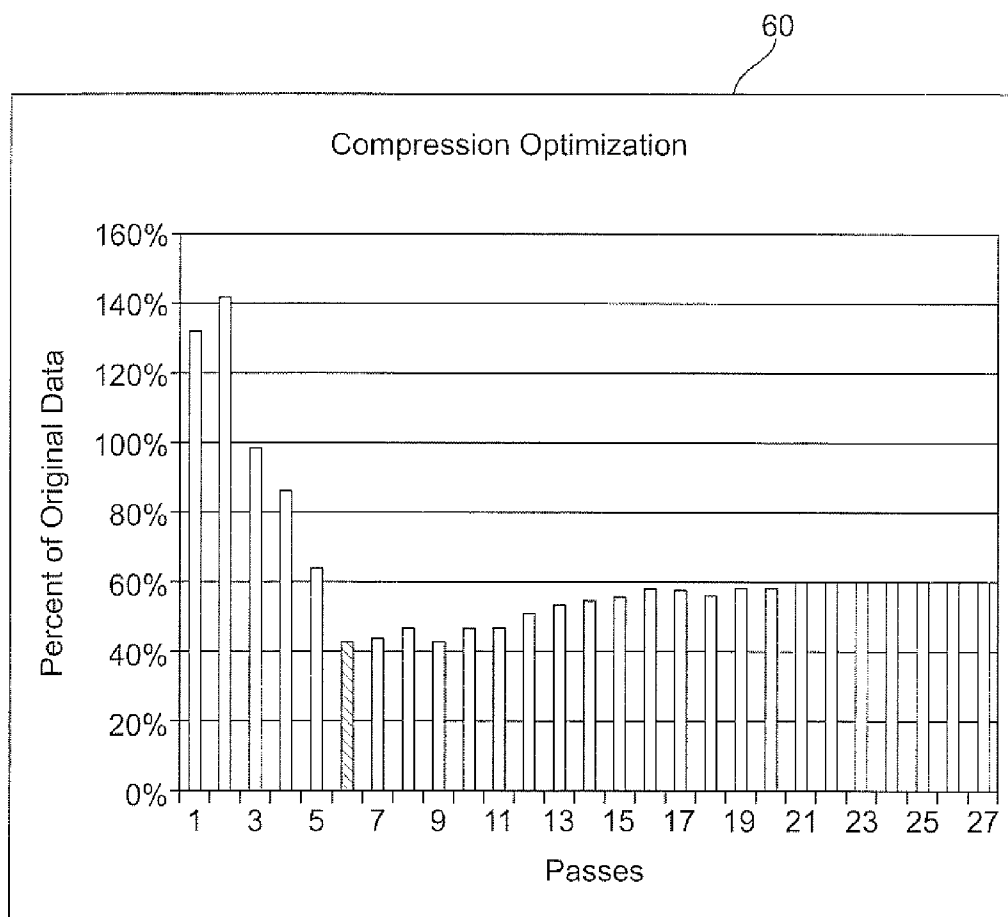


FIG. 61



Symbol	Occurrences	Huffman Code	Original Bits	Total Original Bits Represented	% of File Represented
7	39	1	01011000	312	81.2
5	9	00	11000	45	11.7
3	8	011	100	24	6.2
2	1	01011	10	2	0.5
0	1	0100	0	1	0.3
EOF	1	01010	N/A	N/A	N/A

FIG. 62

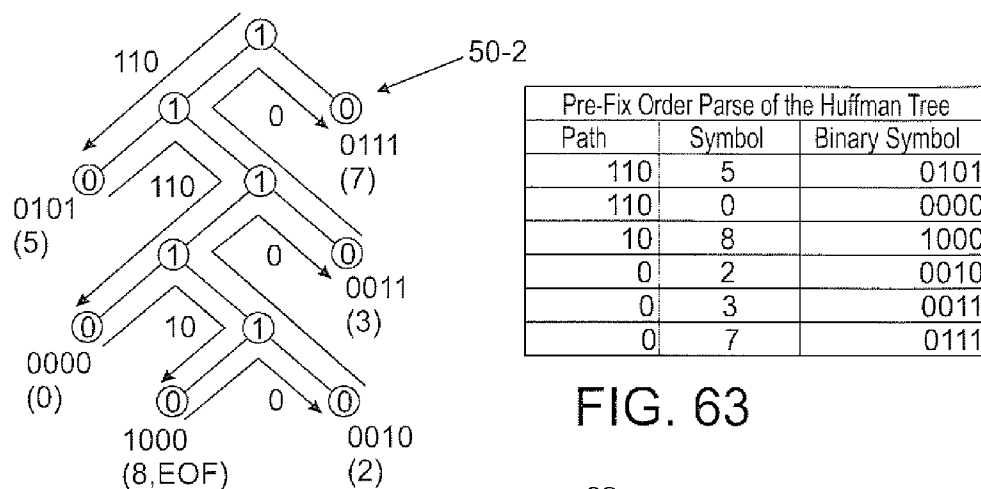


FIG. 63

Binary Powers	Decimal Value	Binary Value	Number of Bits
$2^0$	1	1	1
$2^1$	2	10	2
$2^2$	4	100	3
$2^3$	8	1000	4
$2^4$	16	10000	5

FIG. 64

FIG. 67

FIG. 70

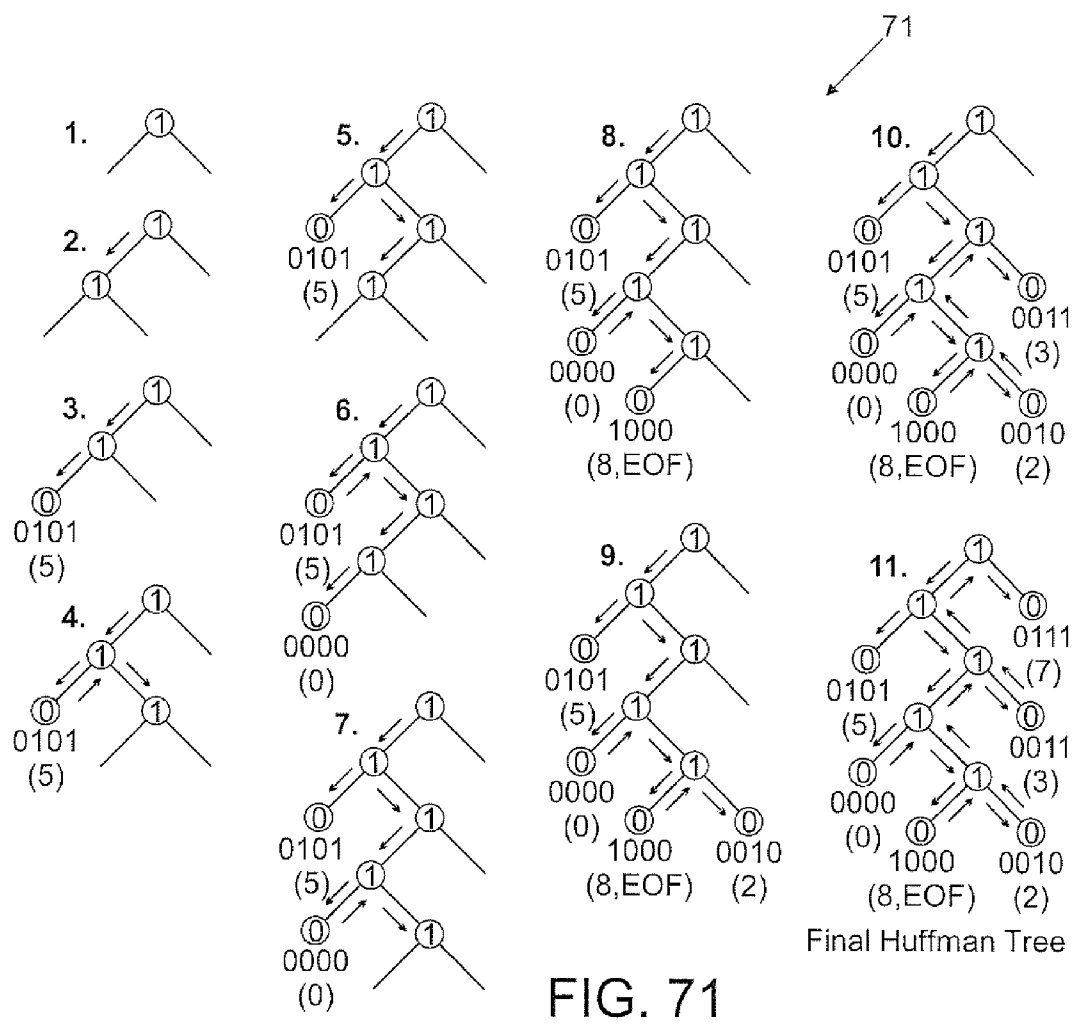


FIG. 71

72

Symbol	Bits per Symbol
2	1
3	2
4	2
5	3
6	3
7	3

FIG. 72

73

Symbol	Tuple	
	First	Last
0	atomic	atomic
1	atomic	atomic
2	1	0
3	2	0
4	1	3
5	4	0
6	2	5
7	0	6

FIG. 73

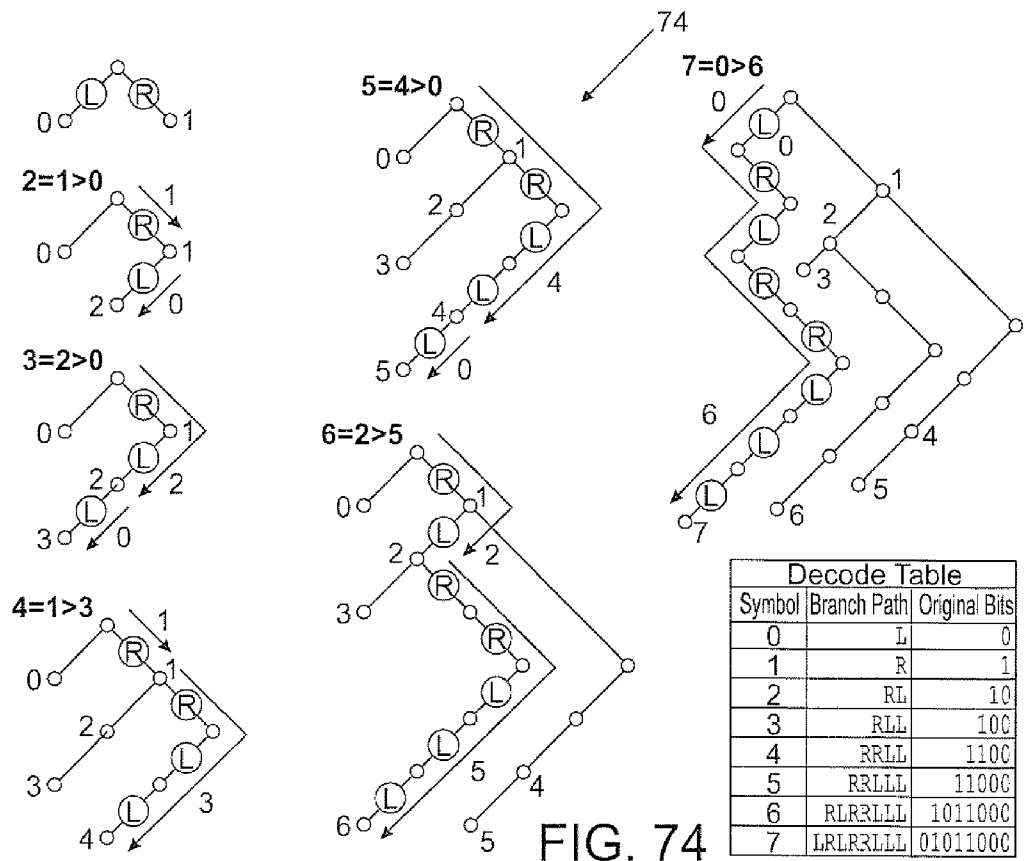


FIG. 74

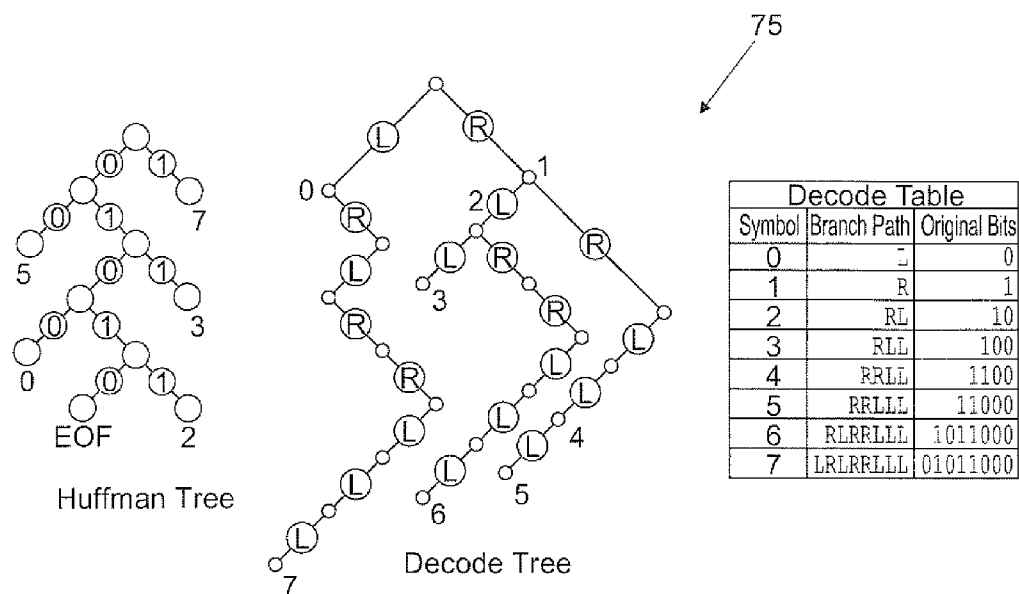


FIG. 75

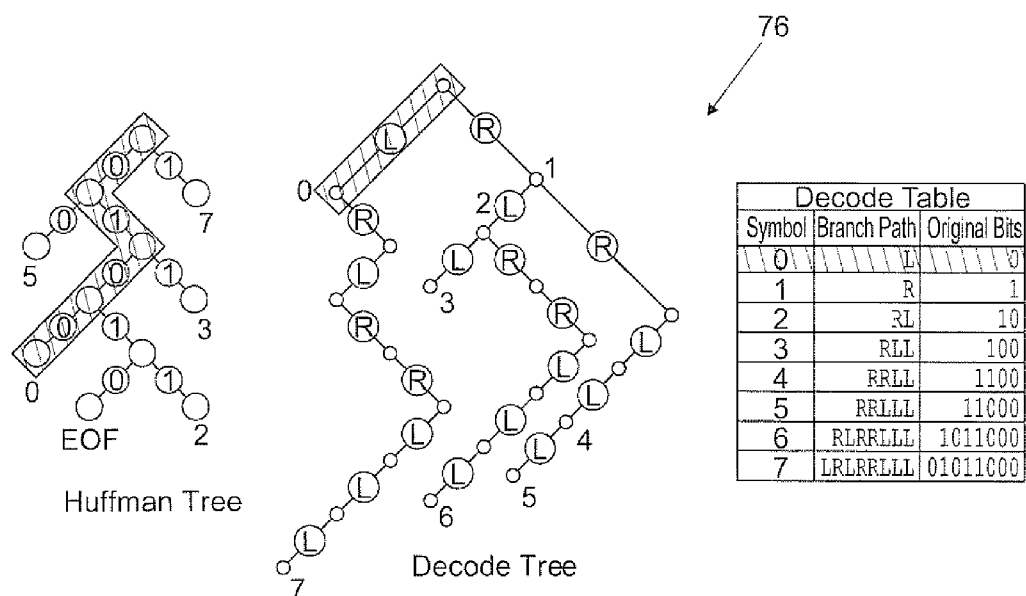


FIG. 76

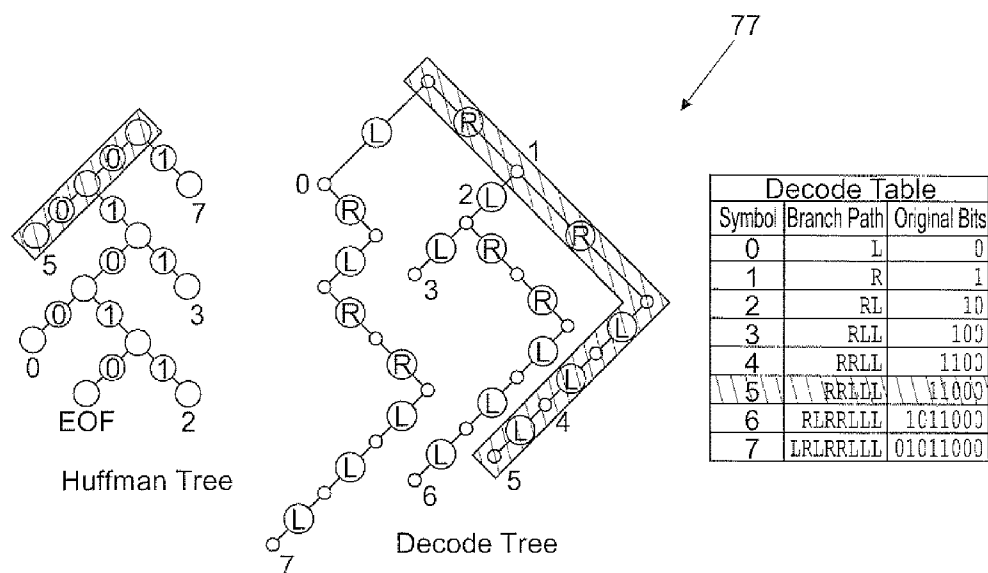


FIG. 77

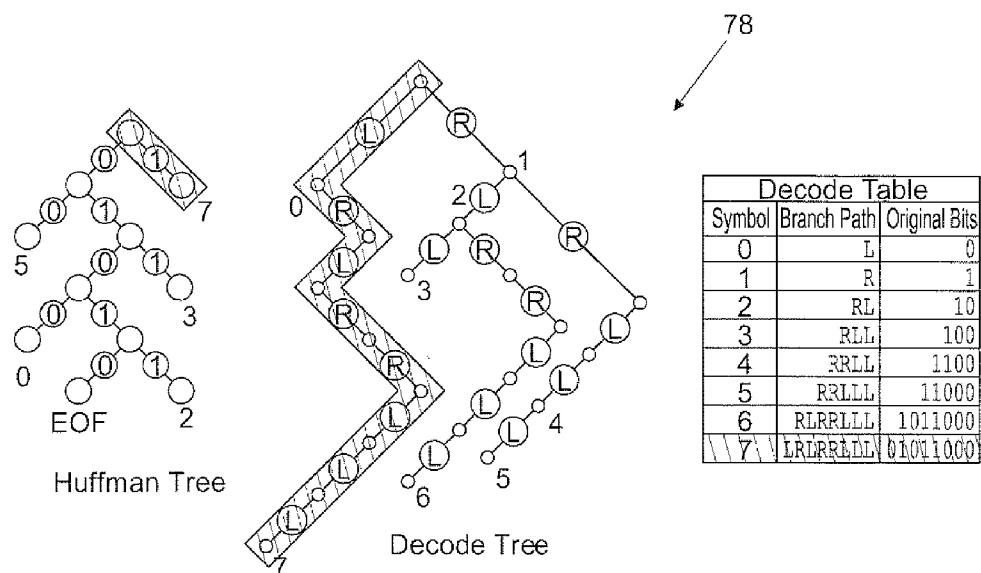


FIG. 78

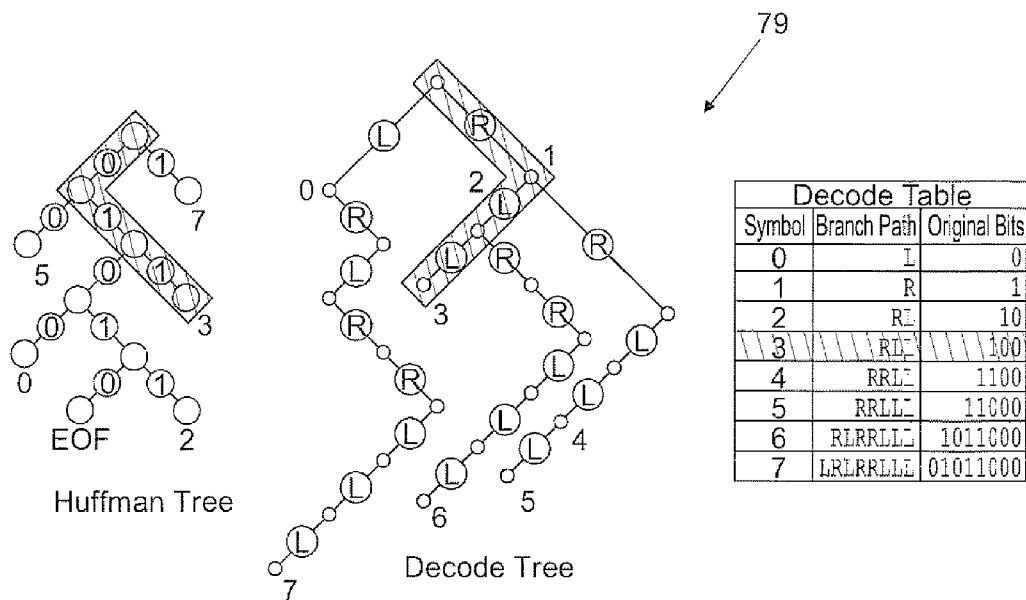


FIG. 79

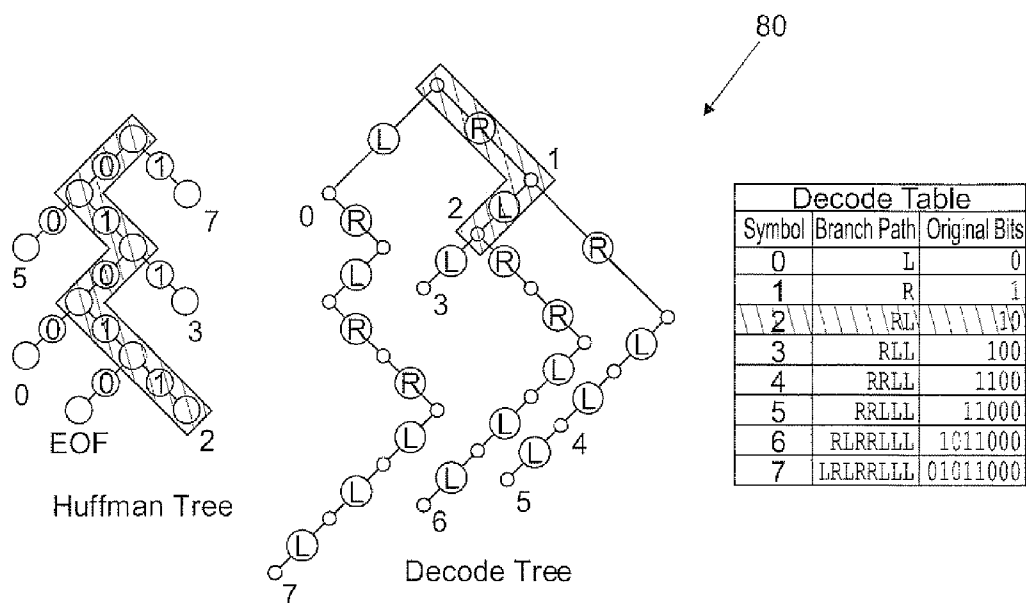


FIG. 80



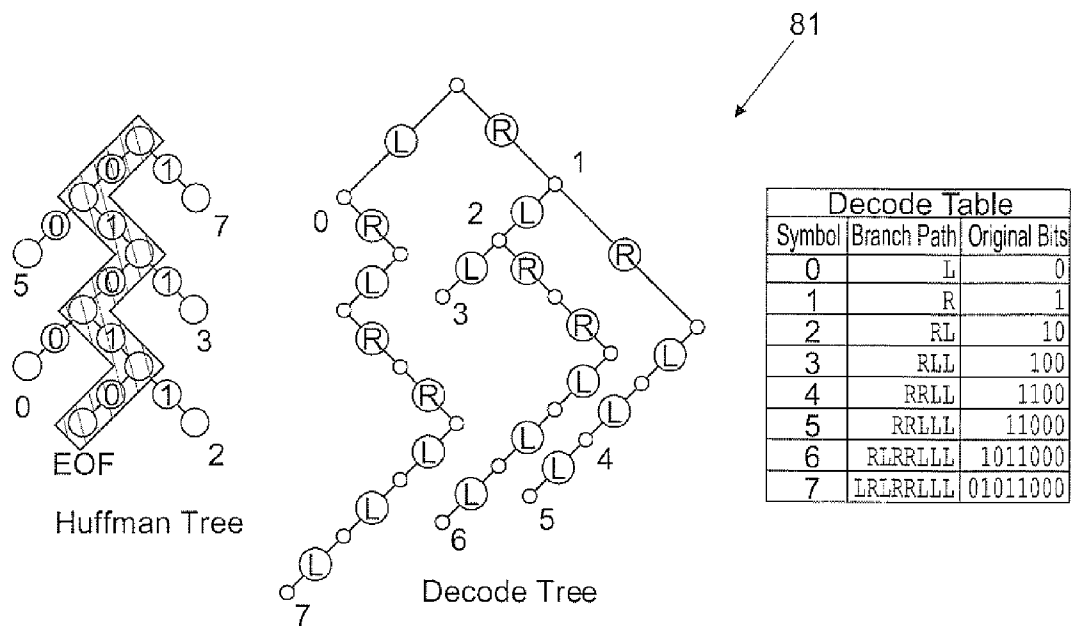


FIG. 81

FIG. 82A

Encoded Data	Symbol and Original Bits Represented
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
011	Symbol 3 = 100
00	Symbol 5 = 11000
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
1	Symbol 7 = 01011000
011	Symbol 3 = 100
00	Symbol 5 = 11000
1	Symbol 7 = 01011000
011	Symbol 3 = 100
00	Symbol 5 = 11000
011	Symbol 3 = 100
00	Symbol 5 = 11000
011	Symbol 3 = 100
00	Symbol 5 = 11000
011	Symbol 3 = 100
00	Symbol 5 = 11000
011	Symbol 3 = 100
00	Symbol 5 = 11000
01011	Symbol 2 = 10
01010	EOF=done

82

FIG. 82B

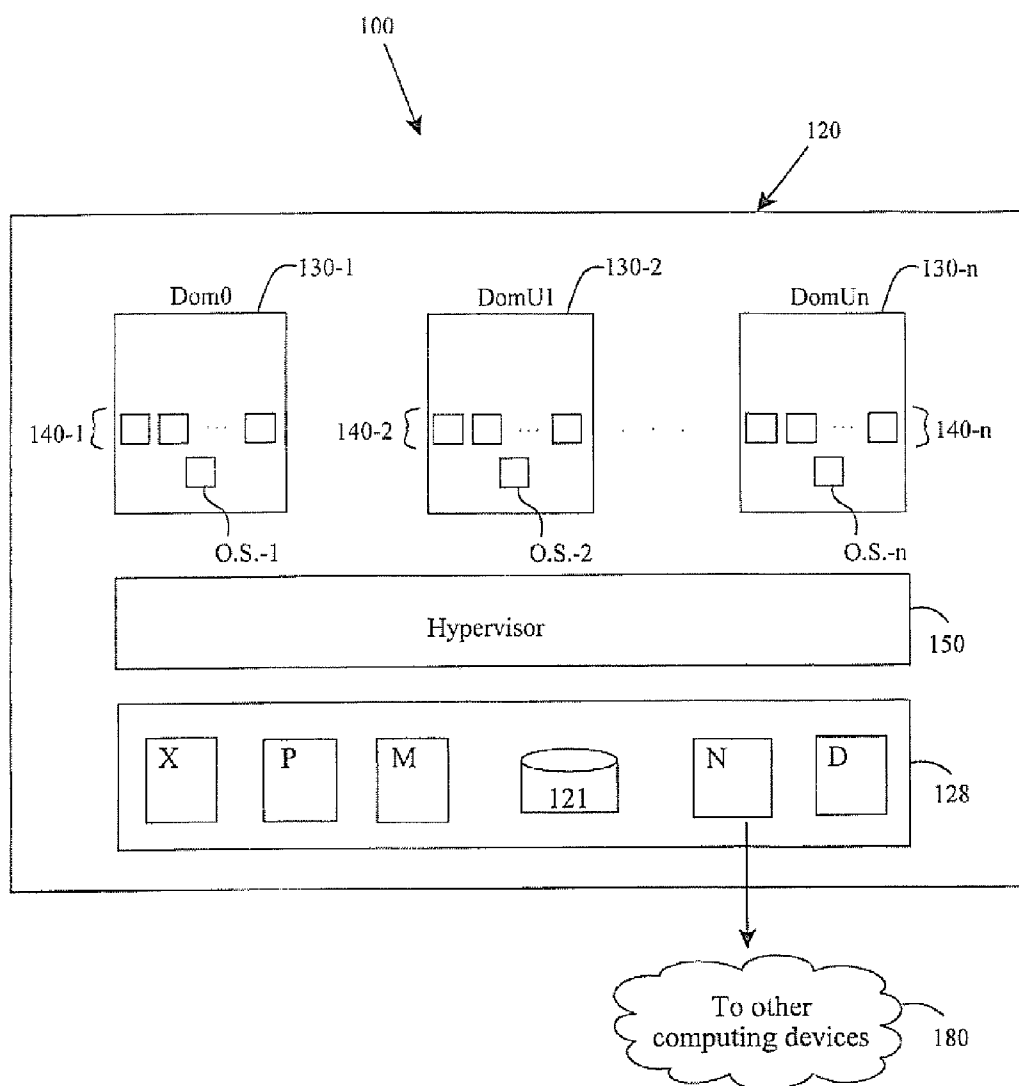


FIG. 83

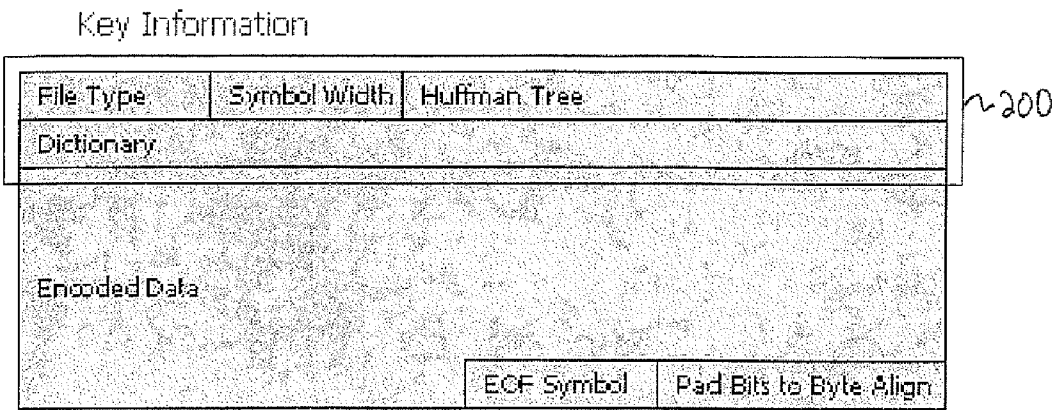
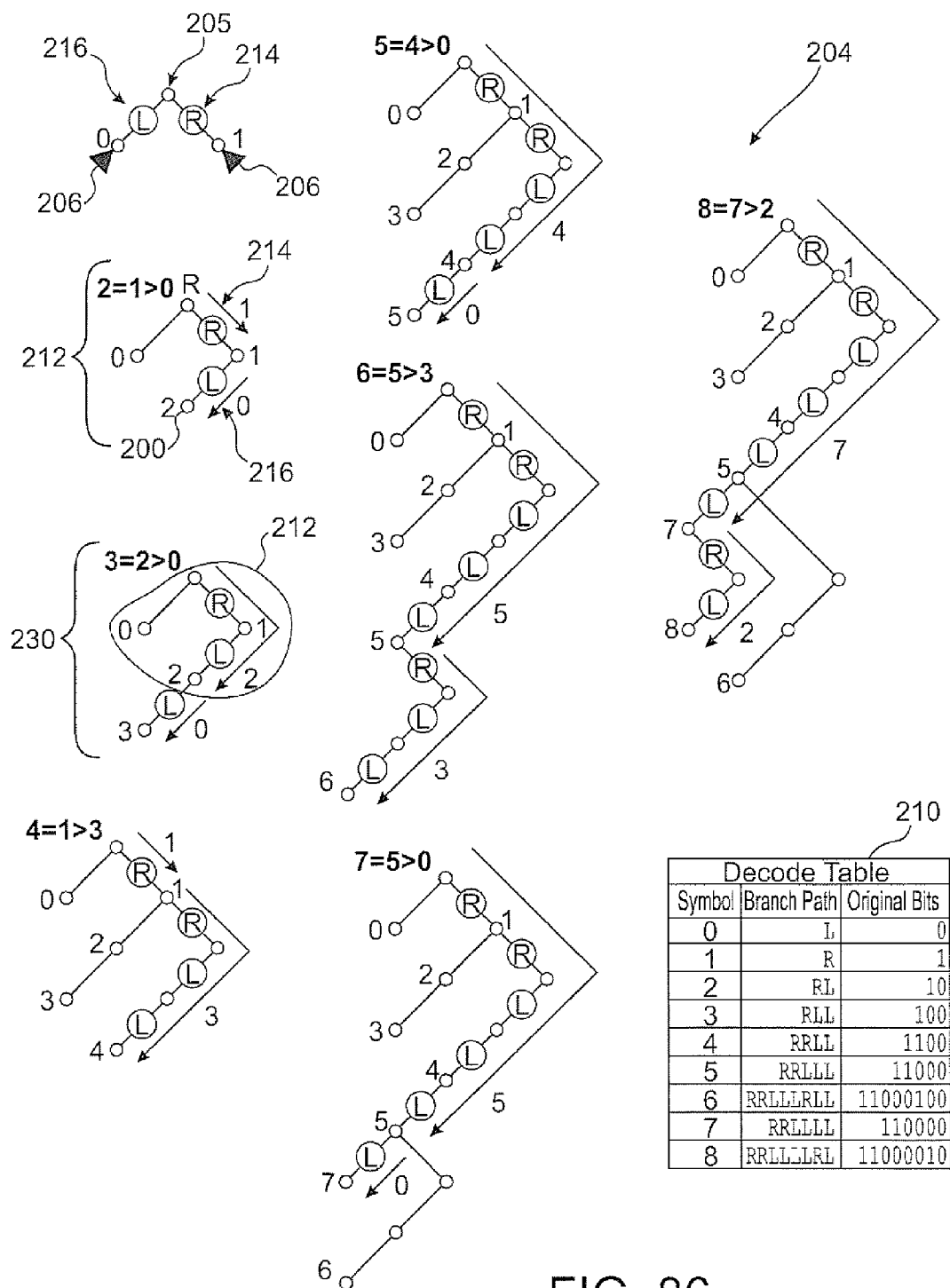


Fig. 84

202

Symbol	Tuple	
	First	Last
0	atomic	atomic
1	atomic	atomic
2	1	0
3	2	0
4	1	3
5	4	0
6	5	3
7	5	0
8	7	2

Fig. 85



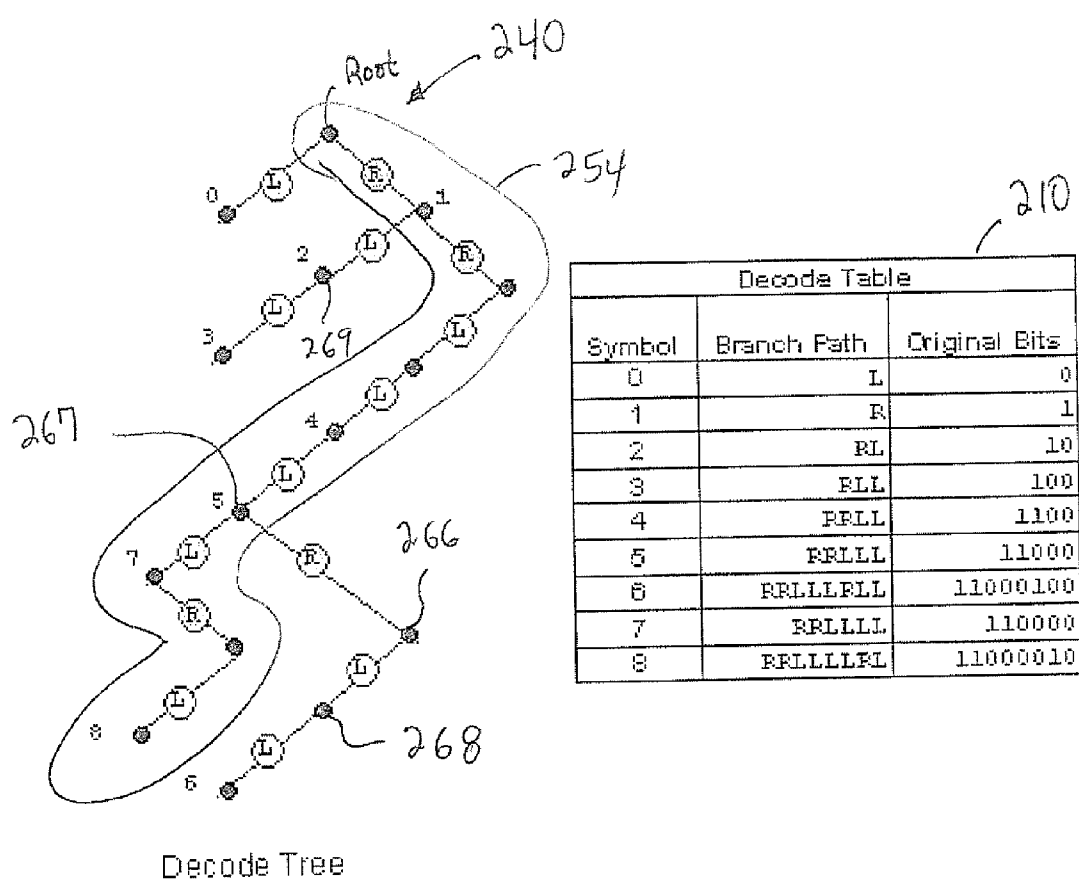


Fig. 87



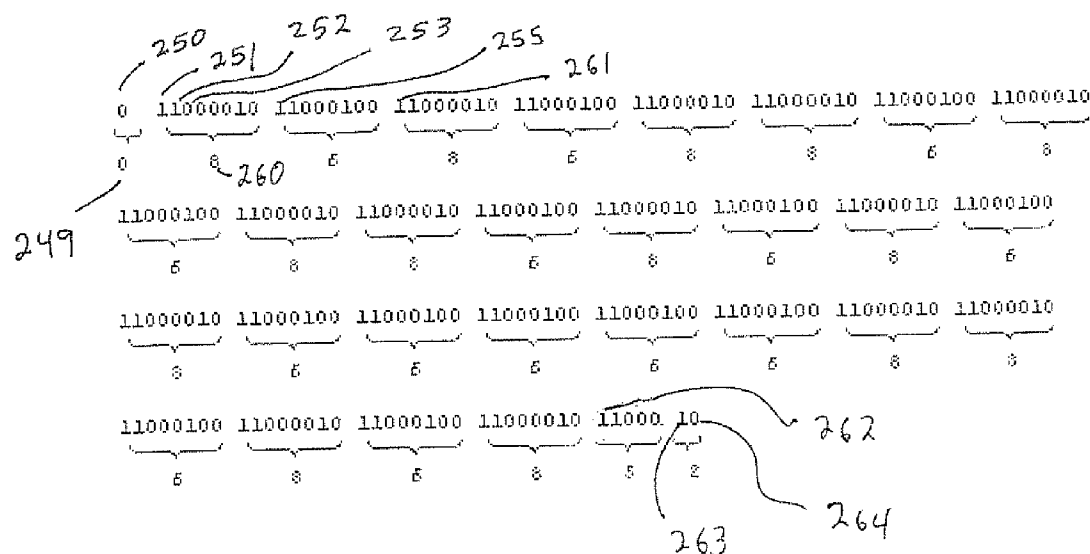


Fig. 88

275

Symbol	Count
0	1
1	0
2	1
3	0
4	0
5	1
6	14
7	0
8	14
EOF	1

Fig. 89

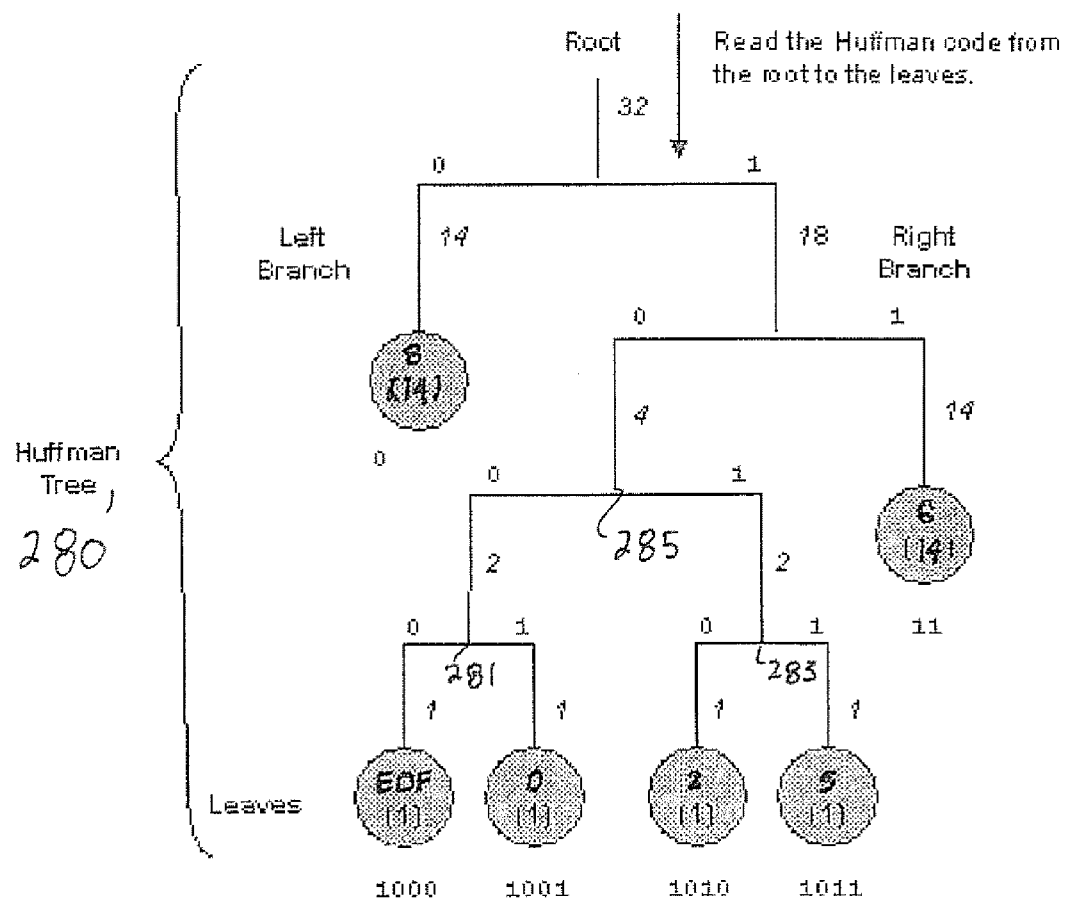
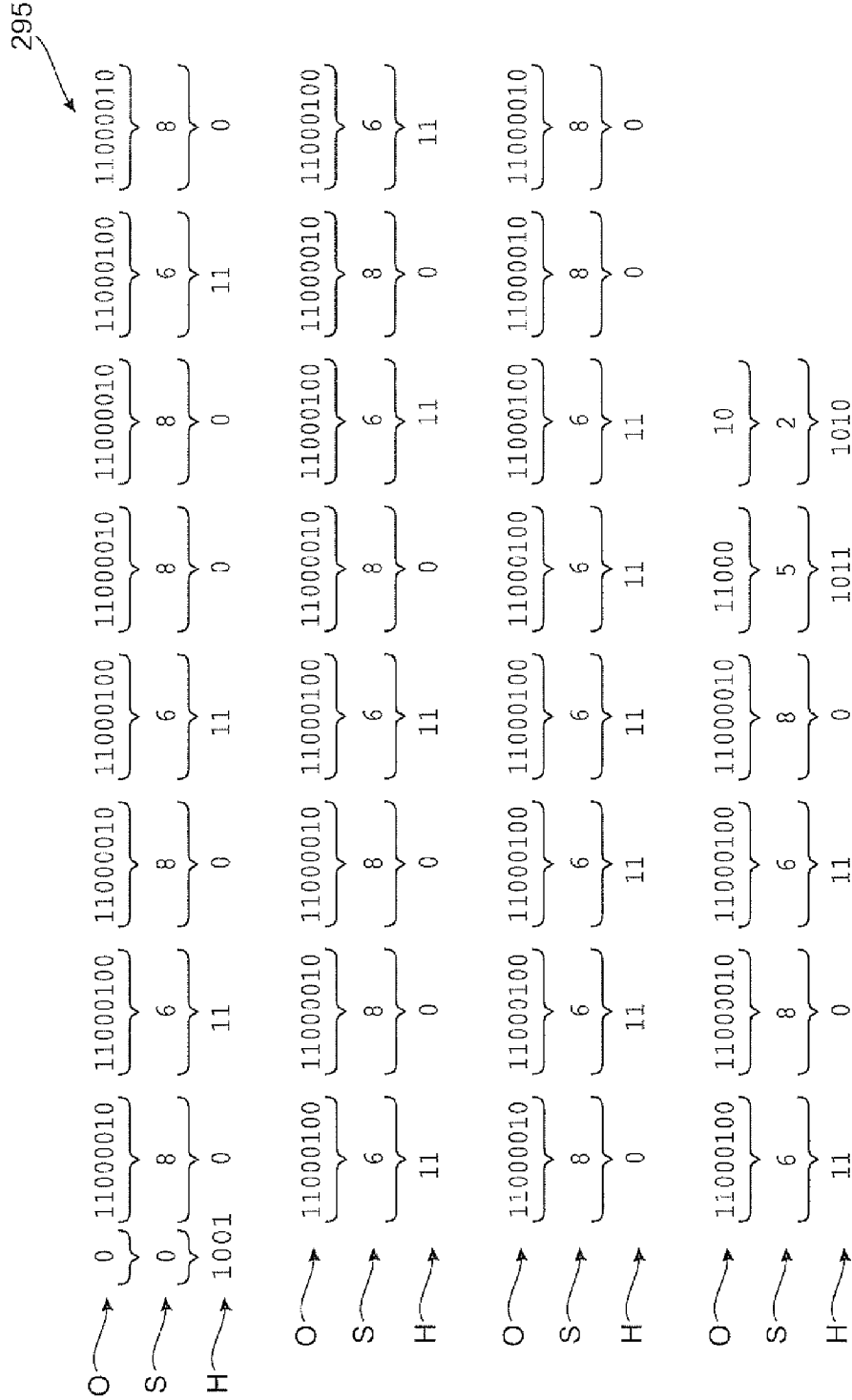


Fig. 90

290

Symbol	Symbol Count	Huffman Code
0	1	1001
1	0	—
2	1	1010
3	0	—
4	0	—
5	1	1011
6	14	11
7	0	—
8	14	0
EOF	1	1000

Fig. 91



Original Bit Stream (O)  
New Bit Stream (S)  
Huffman Coded Bits (H)

FIG. 92

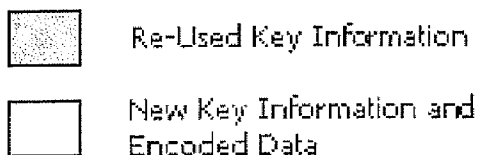
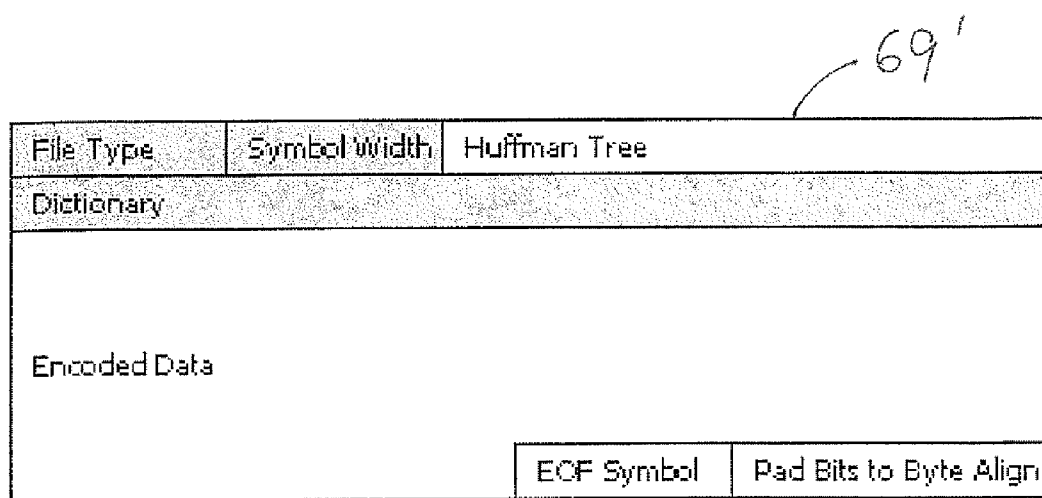
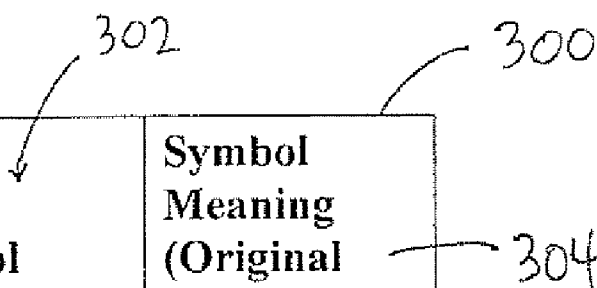


Fig. 93

Term	Definition
Digital spectrum	Information about a file, based on its content, that identifies the file and its position in an N-dimensional universe.
Characteristic digital spectrum	The digital spectrum for a file's data stream. The symbol dictionary from this process defines the N-dimensional space.
Related digital spectrum	The digital spectrum for a related file's data stream as determined by a "fast approximation" process that identifies the file and its position in the same N-dimensional universe as the characteristic digital spectrum.
Azimuth of the symbol frequency vector	A measure of the azimuth of the frequency vector from the origin in N-dimensional space, as measured by applying trigonometry.
Magnitude of the symbol frequency vector	A measure of the distance from the origin in N-dimensional space to the terminal point of the symbol frequency vector, as measured by applying Pythagorean geometry.
Similarity	A measure of the difference in magnitude of the frequency vectors for two digital spectra in an N-dimensional space.
Adjacency	A measure of the distance between two frequency vectors in an N-dimensional space.

**FIG. 94**



Symbol	Symbol Definition	Symbol Meaning (Original Bits)
0	0	0
1	0	1
2	1 > 0	10
3	2 > 0	100
4	1 > 3	1100
5	4 > 0	11000
6	2 > 5	1011000
7	0 > 6	01011000

**FIG. 95**



310

<u>320</u> Symbol	<u>302</u> Symbol Definition	Symbol Meaning (Original Bits) <u>304</u>	<u>312</u> Symbol Count	<u>314</u> Bit Length	<u>316</u> Bit Count
0	0	0	1	1	1
1	0	1	0	1	0
2	1 > 0	10	1	2	2
3	2 > 0	100	8	3	24
4	1 > 3	1100	0	4	0
5	4 > 0	11000	9	5	45
6	2 > 5	1011000	0	7	0
7	0 > 6	01011000	39	8	312
Source File Length (in bits)					384

325

FIG. 96

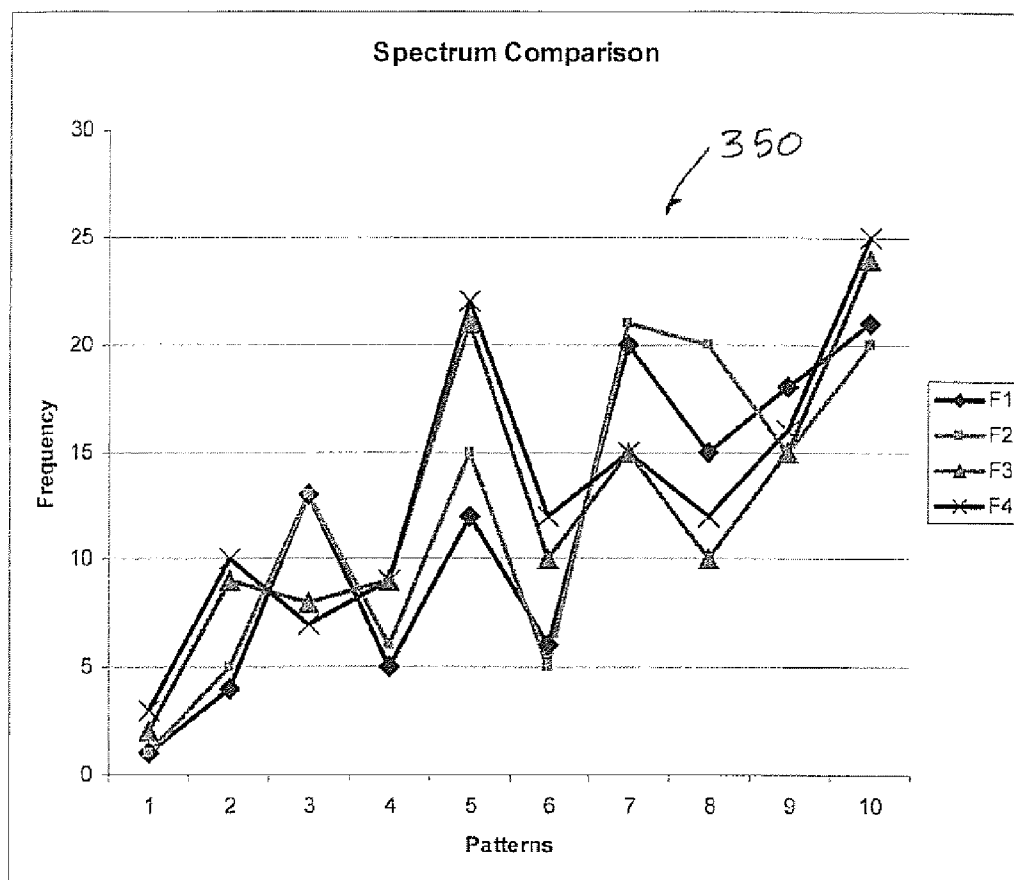


FIG. 97A

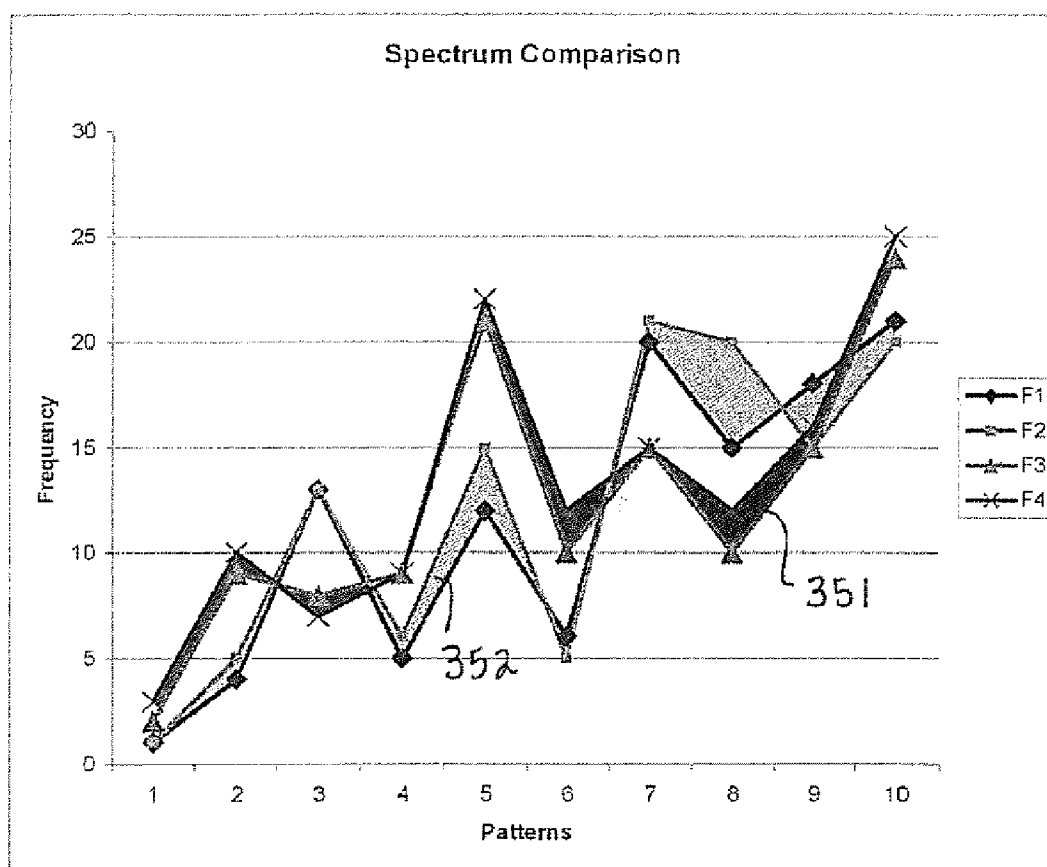


FIG. 97B

## DIGITAL SPECTRUM OF FILE BASED ON CONTENTS

**[0001]** This utility application claims priority to U.S. Provisional Application Ser. Nos. 61/236,571 and 61/271,079, filed Aug. 25, 2009, and Jul. 16, 2009, respectively. Their contents are expressly incorporated herein as if set forth herein.

### FIELD OF THE INVENTION

**[0002]** The present invention relates generally to compression/decompression of data. More particularly, it relates to defining a digital spectrum of a compressed file in order to determine properties that can be compared to other files to ascertain file similarity, adjacency and grouping, to name a few. Vectors and scalar values, among other things, are described for the digital spectrum.

### BACKGROUND OF THE INVENTION

**[0003]** Recent data suggests that nearly eighty-five percent of all data is found in computing files and growing annually at around sixty percent. One reason for the growth is that regulatory compliance acts, statutes, etc., (e.g., Sarbanes-Oxley, HIPAA, PCI) force companies to keep file data in an accessible state for extended periods of time. However, block level operations in computers are too lowly to apply any meaningful interpretation of this stored data beyond taking snapshots and block de-duplication. While other business intelligence products have been introduced to provide capabilities greater than block-level operations, they have been generally limited to structured database analysis. They are much less meaningful when acting upon data stored in unstructured environments.

**[0004]** Unfortunately, entities the world over have paid enormous sums of money to create and store their data, but cannot find much of it later in instances where it is haphazardly arranged or arranged less than intuitively. Not only would locating this information bring back value, but being able to observe patterns in it might also prove valuable despite its usefulness being presently unknown. However, entities cannot expend so much time and effort in finding this data that it outweighs its usefulness. Notwithstanding this, there are still other scenarios, such as government compliance, litigation, audits, etc., that dictate certain data/information be found and produced, regardless of its cost in time, money and effort. Thus, a clear need is identified in the art to better find, organize and identify digital data, especially data left in unstructured states.

**[0005]** In search engine technology, large amounts of unrelated and unstructured digital data can be quickly gathered. However, most engines do little to organize the data other than give a hierarchical presentation. Also, when the engine finds duplicate versions of data, it offers few to no options on eliminating the replication or migrating/relocating redundancies. Thus, a further need in the art exists to overcome the drawbacks of search engines.

**[0006]** When it comes to large amounts of data, whether structured or not, compression techniques have been devised to preserve storage capacity, reduce bandwidth during transmission, etc. With modern compression algorithms, however, they simply exist to scrunch large blocks of data into smaller blocks according to their advertised compression ratios. As is

known, some do it without data loss (lossless) while others do it "lossy." None do it, unfortunately, with a view toward recognizing similarities in the data itself.

**[0007]** From biology, it is known that highly similar species have highly similar DNA strings. In the computing context, consider two word processing files relating to stored baseball statistics. In a first file, words might appear for a baseball batter, such as "batting average," "on base percentage," and "slugging percentage," while a second file might have words for a baseball pitcher, such as "strikeouts," "walks," and "earned runs." Conversely, a third file wholly unrelated to baseball, statistics or sports, may have words such as "environmental protection," "furniture," or whatever comes to mind. It would be exceptionally useful if, during times of compression, or upon later manipulation by an algorithm if "mapping" could recognize the similarity in subject matter in the first two files, although not exact to one another, and provide options to a user. Appreciating that the "words" in the example files are represented in the computing context as binary bits (1's or 0's), which occurs by converting the English alphabet into a series of 1's and 0's through application of ASCII encoding techniques, it would be further useful if the compression algorithm could first recognize the similarity in subject matter of the first two files at the level of raw bit data. The reason for this is that not all files have words and instead might represent pictures (e.g., jpeg) or spread sheets of numbers.

**[0008]** Appreciating that certain products already exist in the above-identified market space, clarity on the need in the art is as follows. One, present day "keyword matching" is limited to select set of words that have been pulled from a document into an index for matching to the same exact words elsewhere. Two, "Grep" is a modern day technique that searches one or more input files for lines containing an identical match to a specified pattern. Three, "Beyond Compare," and similar algorithms, are line-by-line comparisons of multiple documents that highlight differences between them. Four, block level data de-duplication has no application in compliance contexts, data relocation, or business intelligence.

**[0009]** The need in the art, on the other hand, needs to serve advanced notions of identifying new business intelligence, conducting operations on completely unstructured or haphazard data, and organizing it, providing new useful options to users, providing new user views, providing new encryption products, and identifying highly similar data, to name a few. As a byproduct, solving this need will create new opportunities in minimizing transmission bandwidth and storage capacity, among other things. Naturally, any improvements along such lines should contemplate good engineering practices, such as stability, ease of implementation, unobtrusiveness, etc.

### SUMMARY OF THE INVENTION

**[0010]** Applying the principles and teachings associated with a file's digital spectrum solves the foregoing and other problems. Broadly, methods and apparatus of a digital spectrum is used to compute and communicate a file's informational characteristics. Two representative methods are presented. In the first, a file's informational position may be represented as a vector in an N-dimensional space, where each dimension is defined by a symbol described in the digital spectrum. The position along the axis of any given dimension is described by the frequency (or other derivative informa-

tion) of occurrence of that symbol. Relative to the origin of the N-dimensional space, the file's informational position can be computed. Comparing positions reveals similarity, or not, of the files.

[0011] In another method, the digital spectrum defines a line graph, wherein each symbol and its frequency of occurrence define a point on the line. A distance function between two spectra line graphs is computed. Comparing the values derived from the distance function reveals similarity, or not, of the files. Also, total numbers of bits in the files are extracted by knowing the lengths of the original bits corresponding to every symbol. A symbol bit length spectrum is defined. Further derivative comparison functions are anticipated using the symbol bit length spectrum and file length.

[0012] Executable instructions loaded on one or more computing devices for undertaking the foregoing are also contemplated as are computer program products available as a download or on a computer readable medium. The computer program products are also available for installation on a network appliance or an individual computing device.

[0013] These and other embodiments of the present invention will be set forth in the description which follows, and in part will become apparent to those of ordinary skill in the art by reference to the following description of the invention and referenced drawings or by practice of the invention. The claims, however, indicate the particularities of the invention.

#### BRIEF DESCRIPTION OF THE DRAWINGS

[0014] The accompanying drawings incorporated in and forming a part of the specification, illustrate several aspects of the present invention, and together with the description serve to explain the principles of the invention. In the drawings:

[0015] FIG. 1 is a table in accordance with the present invention showing terminology;

[0016] FIG. 2 a table in accordance with the present invention showing a tuple array and tuple nomenclature;

[0017] FIG. 3 is a table in accordance with the present invention showing the counting of tuples in a data stream;

[0018] FIG. 4 is a table in accordance with the present invention showing the Count from FIG. 3 in array form;

[0019] FIG. 5 is Pythagorean's Theorem for use in resolving ties in the counts of highest occurring tuples;

[0020] FIG. 6 is a table in accordance with the present invention showing a representative resolution of a tie in the counts of three highest occurring tuples using Pythagorean's Theorem;

[0021] FIG. 7 is a table in accordance with the present invention showing an alternative resolution of a tie in the counts of highest occurring tuples;

[0022] FIG. 8 is an initial dictionary in accordance with the present invention for the data stream of FIG. 9;

[0023] FIGS. 8-60 are iterative data streams and tables in accordance with the present invention depicting dictionaries, arrays, tuple counts, encoding, and the like illustrative of multiple passes through the compression algorithm;

[0024] FIG. 61 is a chart in accordance with the present invention showing compression optimization;

[0025] FIG. 62 is a table in accordance with the present invention showing compression statistics;

[0026] FIGS. 63-69 are diagrams and tables in accordance with the present invention relating to storage of a compressed file;

[0027] FIGS. 70-82b are data streams, tree diagrams and tables in accordance with the present invention relating to decompression of a compressed file;

[0028] FIG. 83 is a diagram in accordance with the present invention showing a representative computing device for practicing all or some the foregoing;

[0029] FIGS. 84-93 are diagrams in accordance with a "fast approximation" embodiment of the invention that utilizes key information of an earlier compressed file for a file under present consideration having patterns substantially similar to the earlier compressed file; and

[0030] FIGS. 94-97A-B are definitions and diagrams in accordance with the present invention showing a "digital spectrum" embodiment of an encoded file.

#### DETAILED DESCRIPTION OF THE ILLUSTRATED EMBODIMENTS

[0031] In the following detailed description of the illustrated embodiments, reference is made to the accompanying drawings that form a part hereof, and in which is shown by way of illustration, specific embodiments in which the invention may be practiced. These embodiments are described in sufficient detail to enable those skilled in the art to practice the invention and like numerals represent like details in the various figures. Also, it is to be understood that other embodiments may be utilized and that process, mechanical, electrical, arrangement, software and/or other changes may be made without departing from the scope of the present invention. In accordance with the present invention, methods and apparatus are hereinafter described for optimizing data compression of digital data.

[0032] In a representative embodiment, compression occurs by finding highly occurring patterns in data streams, and replacing them with newly defined symbols that require less space to store than the original patterns. The goal is to eliminate as much redundancy from the digital data as possible. The end result has been shown by the inventor to achieve greater compression ratios on certain tested files than algorithms heretofore known.

[0033] In information theory, it is well understood that collections of data contain significant amounts of redundant information. Some redundancies are easily recognized, while others are difficult to observe. A familiar example of redundancy in the English language is the ordered pair of letters QU. When Q appears in written text, the reader anticipates and expects the letter U to follow, such as in the words queen, quick, acquit, and square. The letter U is mostly redundant information when it follows Q. Replacing a recurring pattern of adjacent characters with a single symbol can reduce the amount of space that it takes to store that information. For example, the ordered pair of letters QU can be replaced with a single memorable symbol when the text is stored. For this example, the small Greek letter alpha ( $\alpha$ ) is selected as the symbol, but any could be chosen that does not otherwise appear in the text under consideration. The resultant compressed text is one letter shorter for each occurrence of QU that is replaced with the single symbol ( $\alpha$ ), e.g., "aeen," "cick," "accit," and "scare." Such is also stored with a definition of the symbol alpha ( $\alpha$ ) in order to enable the original data to be restored. Later, the compressed text can be expanded by replacing the symbol with the original letters

QU. There is no information loss. Also, this process can be repeated many times over to achieve further compression.

#### DEFINITIONS

**[0034]** With reference to FIG. 1, a table 10 is used to define terminology used in the below compression method and procedure.

#### Discussion

**[0035]** Redundancy is the superfluous repetition of information. As demonstrated in the QU example above, adjacent characters in written text often form expected patterns that are easily detected. In contrast, digital data is stored as a series of bits where each bit can have only one of two values: off (represented as a zero (0)) and on (represented as a one (1)). Redundancies in digital data, such as long sequences of zeros or ones, are easily seen with the human eye. However, patterns are not obvious in highly complex digital data. The invention's methods and procedures identify these redundancies in stored information so that even highly complex data can be compressed. In turn, the techniques can be used to reduce, optimize, or eliminate redundancy by substituting the redundant information with symbols that take less space to store than the original information. When it is used to eliminate redundancy, the method might originally return compressed data that is larger than the original. This can occur because information about the symbols and how the symbols are encoded for storage must also be stored so that the data can be decompressed later. For example, compression of the word "queen" above resulted in the compressed word "αeen." But a dictionary having the relationship  $QU=\alpha$  also needed to be stored with the word "αeen," which makes a "first pass" through the compression technique increase in size, not decrease. Eventually, however, further "passes" will stop increasing and decrease so rapidly, despite the presence of an ever-growing dictionary size, that compression ratios will be shown to greatly advance the state of the art. By automating the techniques with computer processors and computing software, compression will also occur exceptionally rapidly. In addition, the techniques herein will be shown to losslessly compress the data.

#### The Compression Procedure

**[0036]** The following compression method iteratively substitutes symbols for highly occurring tuples in a data stream. An example of this process is provided later in the document.

#### Prerequisites

**[0037]** The compression procedure will be performed on digital data. Each stored bit has a value of binary 0 or binary 1. This series of bits is referred to as the original digital data.

#### Preparing the Data

**[0038]** The original digital data is examined at the bit level. The series of bits is conceptually converted to a stream of characters, referred to as the data stream that represents the original data. The symbols 0 and 1 are used to represent the respective raw bit values in the new data stream. These symbols are considered to be atomic because all subsequently defined symbols represent tuples that are based on 0 and 1.

**[0039]** A dictionary is used to document the alphabet of symbols that are used in the data stream. Initially, the alphabet consists solely of the symbols 0 and 1.

#### Compressing the Data Stream

**[0040]** The following tasks are performed iteratively on the data stream:

**[0041]** Identifying all possible tuples that can occur for the set of characters that are in the current data stream.

**[0042]** Determining which of the possible tuples occurs most frequently in the current data stream. In the case of a tie, use the most complex tuple. (Complexity is discussed below.)

**[0043]** Creating a new symbol for the most highly occurring tuple, and add it to the dictionary.

**[0044]** Replacing all occurrences of the most highly occurring tuple with the new symbol.

**[0045]** Encoding the symbols in the data stream by using an encoding scheme, such as a path-weighted Huffman coding scheme.

**[0046]** Calculating the compressed file size.

**[0047]** Determining whether the compression goal has been achieved.

**[0048]** Repeating for as long as necessary to achieve optimal compression. That is, if a stream of data were compressed so completely that it was represented by a single bit, it and its complementary dictionary would be larger than the original representation of the stream of data absent the compression. (For example, in the QU example above, if "α" represented the entire word "queen," the word "queen" could be reduced to one symbol, e.g., "α." However, this one symbol and its dictionary (reciting "queen=α" is larger than the original content "queen.") Thus, optimal compression herein recognizes a point of marginal return whereby the dictionary grows too large relative to the amount of compression being achieved by the technique.

Each of these steps is described in more detail below.

#### Identifying all Possible Triples

**[0049]** From FIG. 1, a "tuple" is an ordered pair of adjoining characters in a data stream. To identify all possible tuples in a given data stream, the characters in the current alphabet are systematically combined to form ordered pairs of symbols. The left symbol in the pair is referred to as the "first" character, while the right symbol is referred to as the "last" character. In a larger context, the tuples represent the "patterns" examined in a data stream that will yield further advantage in the art.

**[0050]** In the following example and with any data stream of digital data that can be compressed according to the techniques herein, two symbols (0 and 1) occur in the alphabet and are possibly the only symbols in the entire data stream. By examining them as "tuples," the combination of the 0 and 1 as ordered pairs of adjoining characters reveals only four possible outcomes, i.e., a tuple represented by "00," a tuple represented by "01," a tuple represented by "10," and a tuple represented by "11."

**[0051]** With reference to FIG. 2, these four possibilities are seen in table 12. In detail, the table shows the tuple array for characters 0 and 1. In the cell for column 0 and row 0, the tuple is the ordered pair of 0 followed by 0. The shorthand notation of the tuple in the first cell is "0>0". In the cell for column 0

and row 1, the tuple is 0 followed by 1, or “0>1”. In the cell for column 1 and row 0, the tuple is “1>0”. In the cell for column 1 and row 1, the tuple is “1>1”.

#### Determining the Most Highly Occurring Tuple

**[0052]** With FIG. 2 in mind, it is determined which tuple in a bit stream is the most highly occurring. To do this, simple counting occurs. It reveals how many times each of the possible tuples actually occurs. Each pair of adjoining characters is compared to the possible tuples and the count is recorded for the matched tuple.

**[0053]** The process begins by examining the adjacent characters in position one and two of the data stream. Together, the pair of characters forms a tuple. Advance by one character in the stream and examine the characters in positions two and three. By incrementing through the data stream one character at a time, every combination of two adjacent characters in the data stream is examined and tallied against one of the tuples.

**[0054]** Sequences of repeated symbols create a special case that must be considered when tallying tuples. That is, when a symbol is repeated three or more times, skilled artisans might identify instances of a tuple that cannot exist because the symbols in the tuple belong to other instances of the same tuple. The number of actual tuples in this case is the number of times the symbol repeats divided by two.

**[0055]** For example, consider the data stream 14 in table 16 (FIG. 3) having 10 characters shown as “0110000101.” Upon examining the first two characters 01, a tuple is recognized in the form 0 followed by 1 (0>1). Then, increment forward one character and consider the second and third characters 11, which forms the tuple of 1 followed by 1 (1>1). As progression occurs through the data stream, 9 possible tuple combinations are found: 0>1, 1>1, 1>0, 0>0, 0>0, 0>0, 0>1, 1>0, and 0>1 (element 15, FIG. 3). In the sequence of four sequential zeros (at the fourth through seventh character positions in the data stream “0110000101”), three instances of a 0 followed by a 0 (or 0>0) are identified as possible tuples. It is observed that the second instance of the 0>0 tuple (element 17, FIG. 3) cannot be formed because the symbols are used in the 0>0 tuple before and after it, by prescribed rule. Thus, there are only two possible instances in the COUNT 18, FIG. 3, of the 0>0 tuple, not 3. In turn, the most highly occurring tuple counted in this data stream is 0>1, which occurs 3 times (element 19, FIG. 3). Similarly, tuple 1>1 occurs once (element 20, FIG. 3), while tuple 1>0 occurs twice (element 21, FIG. 3).

**[0056]** After the entire data stream has been examined, the final counts for each tuple are compared to determine which tuple occurs most frequently. In tabular form, the 0 followed by a 1 (tuple 0>1) occurs the most and is referenced at element 19 in table 22, FIG. 4.

**[0057]** In the situation of a tie between two or more tuples, skilled artisans must choose between one of the tuples. For this, experimentation has revealed that choosing the tuple that contains the most complex characters usually results in the most efficient compression. If all tuples are equally complex, skilled artisans can choose any one of the tied tuples and define it as the most highly occurring.

**[0058]** The complexity of a tuple is determined by imagining that the symbols form the sides of a right triangle, and the complexity is a measure of the length of the hypotenuse of that triangle. Of course, the hypotenuse is related to the sum of the squares of the sides, as defined by the Pythagorean Theorem, FIG. 5.

**[0059]** The tuple with the longest hypotenuse is considered the most complex tuple, and is the winner in the situation of a tie between the highest numbers of occurring tuples. The reason for this is that less-complex tuples in the situation of a tie are most likely to be resolved in subsequent passes in the decreasing order of their hypotenuse length. Should a tie in hypotenuse length occur, or a tie in complexity, evidence appears to suggest it does not make a difference which tuple is chosen as the most highly occurring.

**[0060]** For example, suppose that tuples 3>7, 4>4 and 1>5 each occur 356 times when counted (in a same pass). To determine the complexity of each tuple, use the tuple symbols as the two sides of a right triangle and calculate the hypotenuse, FIG. 6. In the instance of 3>7, the side of the hypotenuse is the square root of (three squared (9) plus seven squared (49)), or the square root of 58, or 7.6. In the instance of 4>4, the side of the hypotenuse is the square root of (four squared (16) plus four squared (16)), or the square root of 32, or 5.7. Similar, 1>5 calculates as a hypotenuse of 5.1 as seen in table 23 in the Figure. Since the tuple with the largest hypotenuse is the most complex, 3>7's hypotenuse of 7.6 is considered more complex than either of the tuples 4>4 or 1>5.

**[0061]** Skilled artisans can also use the tuple array to visualize the hypotenuse by drawing lines in the columns and rows from the array origin to the tuple entry in the array, as shown in table 24 in FIG. 7. As seen, the longest hypotenuse is labeled 25, so the 3>7 tuple wins the tie, and is designated as the most highly occurring tuple. Hereafter, a new symbol is created to replace the highest occurring tuple (whether occurring the most outright by count or by tie resolution), as seen below. However, based on the complexity rule, it is highly likely that the next passes will replace tuple 4>4 and then tuple 1>5.

#### Creating a Symbol for the Most Highly Occurring Tuple

**[0062]** As before, a symbol stands for the two adjacent characters that form the tuple and skilled artisans select any new symbol they want provided it is not possibly found in the data stream elsewhere. Also, since the symbol and its definition are added to the alphabet, e.g., if “ $\alpha$ =QU,” a dictionary grows by one new symbol in each pass through the data, as will be seen. A good example of a new symbol for use in the invention is a numerical character, sequentially selected, because numbers provide an unlimited source of unique symbols. In addition, reaching an optimized compression goal might take thousands (or even tens of thousands) of passes through the data stream and redundant symbols must be avoided relative to previous passes and future passes.

#### Replacing the Tuple with the New Symbol

**[0063]** Upon examining the data stream to find all occurrences of the highest occurring tuple, skilled artisans simply substitute the newly defined or newly created symbol for each occurrence of that tuple. Intuitively, substituting a single symbol for two characters compresses the data stream by one character for each occurrence of the tuple that is replaced.

#### Encoding the Alphabet

**[0064]** To accomplish this, counting occurs for how many times that each of the symbols in the current alphabet occurs in the data stream. They then use the symbol count to apply an encoding scheme, such as a path-weighted Huffman coding scheme, to the alphabet. Huffman trees should be within the purview of the artisan's skill set.

**[0065]** The encoding assigns bits to each symbol in the current alphabet that actually appears in the data stream. That is, symbols with a count of zero occurrences are not encoded in the tree. Also, symbols might go “extinct” in the data stream as they are entirely consumed by yet more complex symbols, as will be seen. As a result, the Huffman code tree is rebuilt every time a new symbol is added to the dictionary. This means that the Huffman code for a given symbol can change with every pass. The encoded length of the data stream usually decreases with each pass.

#### Calculating the Compressed File Size

**[0066]** The compressed file size is the total amount of space that it takes to store the Huffman-encoded data stream plus the information about the compression, such as information about the file, the dictionary, and the Huffman encoding tree. The compression information must be saved along with other information so that the encoded data can be decompressed later.

**[0067]** To accomplish this, artisans count the number of times that each symbol appears in the data stream. They also count the number of bits in the symbol’s Huffman code to find its bit length. They then multiply the bit length by the symbol count to calculate the total bits needed to store all occurrences of the symbol. This is then repeated for each symbol. Thereafter, the total bit counts for all symbols are added to determine how many bits are needed to store only the compressed data. To determine the compressed file size, add the total bit count for the data to the number of bits required for the related compression information (the dictionary and the symbol-encoding information).

Determining Whether the Compression Goal has been Achieved

**[0068]** Substituting a tuple with a single symbol reduces the total number of characters in a data stream by one for each instance of a tuple that is replaced by a symbol. That is, for each instance, two existing characters are replaced with one new character. In a given pass, each instance of the tuple is replaced by a new symbol. There are three observed results:

**[0069]** The length of the data stream (as measured by how many characters make up the text) decreases by half the number of tuples replaced.

**[0070]** The number of symbols in the alphabet increases by one.

**[0071]** The number of nodes in the Huffman tree increases by two.

**[0072]** By repeating the compression procedure a sufficient number of times, any series of characters can eventually be reduced to a single character. That “super-symbol” character conveys the entire meaning of the original text. However, the information about the symbols and encoding that is used to reach that final symbol is needed to restore the original data later. As the number of total characters in the text decreases with each repetition of the procedure, the number of symbols increases by one. With each new symbol, the size of the dictionary and the size of the Huffman tree increase, while the size of the data decreases relative to the number of instances of the tuple it replaces. It is possible that the information about the symbol takes more space to store than the original data it replaces. In order for the compressed file size to become smaller than the original data stream size, the text size must decrease faster than the size increases for the dictionary and the Huffman encoding information.

**[0073]** The question at hand is then, what is the optimal number of substitutions (new symbols) to make, and how should those substitutions be determined?

**[0074]** For each pass through the data stream, the encoded length of the text decreases, while the size of the dictionary and the Huffman tree increases. It has been observed that the compressed file size will reach a minimal value, and then increase. The increase occurs at some point because so few tuple replacements are done that the decrease in text size no longer outweighs the increase in size of the dictionary and Huffman tree.

**[0075]** The size of the compressed file does not decrease smoothly or steadily downward. As the compression process proceeds, the size might plateau or temporarily increase. In order to determine the true (global) minimum, it is necessary to continue some number of iterations past the each new (local) minimum point. This true minimal value represents the optimal compression for the data stream using this method.

**[0076]** Through experimentation, three conditions have been found that can be used to decide when to terminate the compression procedure: asymptotic reduction, observed low, and single character. Each method is described below. Other terminating conditions might be determined through further experimentation.

#### Asymptotic Reduction

**[0077]** An asymptotic reduction is a concession to processing efficiency, rather than a completion of the procedure. When compressing larger files (100 kilobytes (KB) or greater), after several thousand passes, each additional pass produces only a very small additional compression. The compressed size is still trending downward, but at such a slow rate that additional compute time is not warranted.

**[0078]** Based on experimental results, the process is terminated if at least 1000 passes have been done, and less than 1% of additional data stream compression has occurred in the last 1000 passes. The previously noted minimum is therefore used as the optimum compressed file.

#### Observed Low

**[0079]** A reasonable number of passes have been performed on the data and in the last reasonable number of passes a new minimum encoded file size has not been detected. It appears that further passes only result in a larger encoded file size.

**[0080]** Based on experimental results, the process is terminated if at least 1000 passes have been done, and in the last 10% of the passes, a new low has not been established. The previously noted minimum is then used as the optimum compressed file.

#### Single Character

**[0081]** The data stream has been reduced to exactly one character. This case occurs if the file is made up of data that can easily reduce to a single symbol, such a file filled with a repeating pattern. In cases like this, compression methods other than this one might result in smaller compressed file sizes.

#### How the Procedure Optimizes Compression

**[0082]** The representative embodiment of the invention uses Huffman trees to encode the data stream that has been



progressively shortened by tuple replacement, and balanced against the growth of the resultant Huffman tree and dictionary representation.

**[0083]** The average length of a Huffman encoded symbol depends upon two factors:

**[0084]** How many symbols must be represented in the Huffman tree

**[0085]** The distribution of the frequency of symbol use

**[0086]** The average encoded symbol length grows in a somewhat stepwise fashion as more symbols are added to the dictionary. Because the Huffman tree is a binary tree, increases naturally occur as the number of symbols passes each level of the power of 2 (2, 4, 8, 16, 32, 64, etc.). At these points, the average number of bits needed to represent any given symbol normally increases by 1 bit, even though the number of characters that need to be encoded decreases. Subsequent compression passes usually overcome this temporary jump in encoded data stream length.

**[0087]** The second factor that affects the efficiency of Huffman coding is the distribution of the frequency of symbol use. If one symbol is used significantly more than any other, it can be assigned a shorter encoding representation, which results in a shorter encoded length overall, and results in maximum compression. The more frequently a symbol occurs, the shorter the encoded stream that replaces it. The less frequently a symbol occurs, the longer the encoded stream that replaces it.

**[0088]** If all symbols occur at approximately equal frequencies, the number of symbols has the greater effect than does the size of the encoded data stream. Supporting evidence is that maximum compression occurs when minimum redundancy occurs, that is, when the data appears random. This state of randomness occurs when every symbol occurs at the same frequency as any other symbol, and there is no discernable ordering to the symbols.

**[0089]** The method and procedure described in this document attempt to create a state of randomness in the data stream. By replacing highly occurring tuples with new symbols, eventually the frequency of all symbols present in the data stream becomes roughly equal. Similarly, the frequency of all tuples is also approximately equal. These two criteria (equal occurrence of every symbol and equal occurrence of ordered symbol groupings) is the definition of random data. Random data means no redundancy. No redundancy means maximum compression.

**[0090]** This method and procedure derives optimal compression from a combination of the two factors. It reduces the number of characters in the data stream by creating new symbols to replace highly occurring tuples. The frequency distribution of symbol occurrence in the data stream tends to equalize as oft occurring symbols are eliminated during tuple replacement. This has the effect of flattening the Huffman tree, minimizing average path lengths, and therefore, minimizing encoded data stream length. The number of newly created symbols is held to a minimum by measuring the increase in dictionary size against the decrease in encoded data stream size.

#### Example of Compression

**[0091]** To demonstrate the compression procedure, a small data file contains the following simple ASCII characters:

**[0092]** aaaaaaaaaaaaaaaaaaaaaa-  
baaabaaaaaaaaababbbbbb

**[0093]** Each character is stored as a sequence of eight bits that correlates to the ASCII code assigned to the character. The bit values for each character are:

**[0094]** a=01100001

**[0095]** b=01100010

**[0096]** The digital data that represents the file is the original data that we use for our compression procedure. Later, we want to decompress the compressed file to get back to the original data without data loss.

#### Preparing the Data Stream

**[0097]** The digital data that represents the file is a series of bits, where each bit has a value of 0 or 1. We want to abstract the view of the bits by conceptually replacing them with symbols to form a sequential stream of characters, referred to as a data stream.

**[0098]** For our sample digital data, we create two new symbols called 0 and 1 to represent the raw bit values of 0 and 1, respectively. These two symbols form our initial alphabet, so we place them in the dictionary **26**, FIG. **8**.

**[0099]** The data stream **30** in FIG. **9** represents the original series of bits in the stored file, e.g., the first eight bits **32** are "01100001" and correspond to the first letter "a" in the data file. Similarly, the very last eight bits **34** are "01100010" and correspond to the final letter "b" in the data file, and each of the 1's and 0's come from the ASCII code above.

**[0100]** Also, the characters in data stream **30** are separated with a space for user readability, but the space is not considered, just the characters. The space would not occur in computer memory either.

#### Compressing the Data Stream

**[0101]** The data stream **30** of FIG. **9** is now ready for compression. The procedure will be repeated until the compression goal is achieved. For this example, the compression goal is to minimize the amount of space that it takes to store the digital data.

#### Initial Pass

**[0102]** For the initial pass, the original data stream and alphabet that were created in "Preparing the Data Stream" are obtained.

#### Identifying all Possible Tuples

**[0103]** An easy way to identify all possible combinations of the characters in our current alphabet (at this time having 0 and 1) is to create a tuple array (table **35**, FIG. **10**). Those symbols are placed or fitted as a column and row, and the cells are filled in with the tuple that combines those symbols. The columns and rows are constructed alphabetically from left to right and top to bottom, respectively, according to the order that the symbols appear in our dictionary. For this demonstration, we will consider the symbol in a column to be the first character in the tuple, and the symbol in a row to be the last character in the tuple. To simplify the presentation of tuples in each cell, we will use the earlier-described notation of "first>last" to indicate the order of appearance in the pair of characters, and to make it easier to visually distinguish the symbols in the pair. The tuples shown in each cell now represent the patterns we want to look for in the data stream.

**[0104]** For example, the table **35** shows the tuple array for characters 0 and 1. In the cell for column 0 and row 0, the tuple is the ordered pair of 0 followed by 0. The shorthand notation

of the tuple in the first cell is “0>0”. In the cell for column 0 and row 1, the tuple is 0 followed by 1, or “0>1”. In the cell for column 1 and row 0, the tuple is “1>0”. In the cell for column 1 and row 1, the tuple is “1>1”. (As skilled artisans will appreciate, most initial dictionaries and original tuple arrays will be identical to these. The reason is that computing data streams will all begin with a stream of 1’s and 0’s having two symbols only.)

#### Determining the Highly Occurring Tuple

**[0105]** After completion of the tuple array, we are ready to look for the tuples in the data stream **30**, FIG. **9**. We start at the beginning of the data stream with the first two characters “01” labeled element **37**. We compare this pair of characters to our known tuples, keeping in mind that order matters. We match the pair to a topic, and add one count for that instance. We move forward by one character, and look at the pair of characters **38** in positions two and three in the data stream, or “11.” We compare and match this pair to one of the tuples, and add one count for that instance. We continue tallying occurrences of the tuples in this manner until we reach the end of the data stream. In this instance, the final tuple is “10” labeled **39**. By incrementing through the data stream one character at a time, we have considered every combination of two adjacent characters in the data stream, and tallied each instance against one of the tuples. We also consider the rule for sequences of repeated symbols, described above, to determine the actual number of instances for the tuple that is defined by pairs of that symbol.

**[0106]** For example, the first two characters in our sample data stream are 0 followed by 1. This matches the tuple 0>1, so we count that as one instance of the tuple. We step forward one character. The characters in positions two and three are 1 followed by 1, which matches the tuple 1>1. We count it as one instance of the 1>1 tuple. We consider the sequences of three or more zeros in the data stream (e.g., 01100001 . . . ) to determine the actual number of tuples for the 0>0 tuple. We repeat this process to the end of the data set with the count results in table **40**, FIG. **11**.

**[0107]** Now that we have gathered statistics for how many times each tuple appears in the data stream **30**, we compare the total counts for each tuple to determine which pattern is the most highly occurring. The tuple that occurs most frequently is a tie between a 1 followed by 0 (1>0), which occurs 96 times, and a 0 followed by 1 (0>1), which also occurs 96 times. As discussed above, skilled artisans then choose the most complex tuple and do so according to Pythagorean’s Theorem. The sum of the squares for each tuple is the same, which is 1 (1+0) and 1 (0+1). Because they have the same complexity, it does not matter which one is chosen as the highest occurring. In this example, we will choose tuple 1>0.

**[0108]** We also count the number of instances of each of the symbols in the current alphabet as seen in table **41**, FIG. **12**. The total symbol count in the data stream is 384 total symbols that represent 384 bits in the original data. Also, the symbol 0 appears 240 times in original data stream **30**, FIG. **9**, while the symbol 1 only appears 144 times.

#### Pass 1

**[0109]** In this next pass, we replace the most highly occurring tuple from the previous pass with a new symbol, and then we determine whether we have achieved our compression goal.

#### Creating a Symbol for the Highly Occurring Tuple

**[0110]** We replace the most highly occurring tuple from the previous pass with a new symbol and add it to the alphabet.

Continuing the example, we add a new symbol 2 to the dictionary and define it with the tuple defined as 1 followed by 0 (1>0). It is added to the dictionary **26**’ as seen in FIG. **13**. (Of course, original symbol 0 is still defined as a 0, while original symbol 1 is still defined as a 1. Neither of these represent a first symbol followed a last symbol which is why dashes appear in the dictionary **26**’ under “Last” for each of them.)

#### Replacing the Tuple with the New Symbol

**[0111]** In the original data stream **30**, every instance of the tuple 1>0 is now replaced with the new, single symbol. In our example data stream **30**, FIG. **9**, the 96 instances of the tuple 1>0 have been replaced with the new symbol “2” to create the output data stream **30**’, FIG. **14**, that we will use for this pass. As skilled artisans will observe, replacing ninety-six double instances of symbols with a single, new symbol shrinks or compresses the data stream **30**’ in comparison to the original data stream **30**, FIG. **8**.

#### Encoding the Alphabet

**[0112]** After we compress the data stream by using the new symbol, we use a path-weighted Huffman coding scheme to assign bits to each symbol in the current alphabet.

**[0113]** To do this, we again count the number of instances of each of the symbols in the current alphabet (now having “0,” “1” and “2.”) The total symbol count in the data stream is 288 symbols as seen in table **41**’, FIG. **15**. We also have one end-of-file (EOF) symbol at the end of the data stream (not shown).

**[0114]** Next, we use the counts to build a Huffman binary code tree. 1) List the symbols from highest count to lowest count. 2) Combine the counts for the two least frequently occurring symbols in the dictionary. This creates a node that has the value of the sum of the two counts. 3) Continue combining the two lowest counts in this manner until there is only one symbol remaining. This generates a Huffman binary code tree.

**[0115]** Finally, label the code tree paths with zeros (0s) and ones (1s). The Huffman coding scheme assigns shorter code words to the more frequent symbols, which helps reduce the size length of the encoded data. The Huffman code for a symbol is defined as the string of values associated with each path transition from the root to the symbol terminal node.

**[0116]** With reference to FIG. **16**, the tree **50** demonstrates the process of building the Huffman tree and code for the symbols in the current alphabet. We also create a code for the end of file marker that we placed at the end of the data stream when we counted the tuples. In more detail, the root contemplates 289 total symbols, i.e., the 288 symbols for the alphabet “0,” “1” and “2” plus one EOF symbol. At the leaves, the “0” is shown with its counts 144 the “1” with its count of 48, the “2” with its count of 96 and the EOF with its count of 1. Between the leaves and root, the branches define the count in a manner skilled artisans should readily understand.

**[0117]** In this compression procedure, we will re-build a Huffman code tree every time we add a symbol to the current dictionary. This means that the Huffman code for a given symbol can change with every compression pass.

#### Calculating the Compressed File Size

**[0118]** From the Huffman tree, we use its code to evaluate the amount of space needed to store the compressed data as seen in table **52**, FIG. **17**. First, we count the number of bits in the Huffman code for each symbol to find its bit length **53**.

Next, we multiply a symbol's bit length by its count **54** to calculate the total bits **55** used to store the occurrences of that symbol. We add the total bits **56** needed for all symbols to determine how many bits are needed to store only the compressed data. As seen, the current data stream **30'**. FIG. **14** requires 483 bits to store only the information.

**[0119]** To know whether we achieved optimal compression, we must consider the total amount of space that it takes to store the compressed data plus the information about the compression that we need to store in order to decompress the data later. We also must store information about the file, the dictionary, and the Huffman tree. The table **57** in FIG. **18** shows the total compression overhead as being 25 bits, which brings the compressed size of the data stream to 508 bits, or 483 bits plus 25 bits.

Determining Whether the Compression Goal has been Achieved

**[0120]** Finally, we compare the original number of bits (**384**, FIG. **12**) to the current number of bits (**508**) that are needed for this compression pass. We find that it takes 1.32 times as many bits to store the compressed data as it took to store the original data, table **58**, FIG. **19**. This is not compression at all, but expansion.

**[0121]** In early passes, however, we expect to see that the substitution requires more space than the original data because of the effect of carrying a dictionary, adding symbols, and building a tree. On the other hand, skilled artisans should observe an eventual reduction in the amount of space needed as the compression process continues. Namely, as the size of the data set decreases by the symbol replacement method, the size grows for the symbol dictionary and the Huffman tree information that we need for decompressing the data.

#### Pass 2

**[0122]** In this pass, we replace the most highly occurring tuple from the previous pass (pass 1) with still another new symbol, and then we determine whether we have achieved our compression goal.

#### Identifying all Possible Tuples

**[0123]** As a result of the new symbol, the tuple array is expanded by adding the symbol that was created in the previous pass. Continuing our example, we add 2 as a first symbol and last symbol, and enter the tuples in the new cells of table **35'**, FIG. **20**.

#### Determining the Highly Occurring Tuple

**[0124]** As before, the tuple array identifies the tuples that we look for and tally in our revised alphabet. As seen in table **40'**, FIG. **21**, the Total Symbol Count=288. The tuple that occurs most frequently when counting the data stream **30'**. FIG. **14**, is the character 2 followed by the character 0 (**2>0**). It occurs 56 times as seen circled in table **40'**, FIG. **21**.

#### Creating a Symbol for the Highly Occurring Tuple

**[0125]** We define still another new symbol "3" to represent the most highly occurring tuple **2>0**, and add it to the dictionary **26"**, FIG. **22**, for the alphabet that was developed in the previous passes.

#### Replacing the Tuple with the New Symbol

**[0126]** In the data stream **30'**. FIG. **14**, we replace every instance of the most highly occurring tuple with the new

single symbol. We replace the 56 instances of the **2>0** tuple with the symbol 3 and the resultant data stream **30'''** is seen in FIG. **23**.

#### Encoding the Alphabet

**[0127]** As demonstrated above, we count the number of symbols in the data stream, and use the count to build a Huffman tree and code for the current alphabet. The total symbol count has been reduced from 288 to 234 (e.g.,  $88+48+40+58$ , but not including the EOF marker) as seen in table **41"**, FIG. **24**.

#### Calculating the Compressed File Size

**[0128]** We need to evaluate whether our substitution reduces the amount of space that it takes to store the data. As described above, we calculate the total bits needed (**507**) as in table **52'**. FIG. **25**.

**[0129]** In table **57'**, FIG. **26**, the compression overhead is calculated as 38 bits.

Determining Whether the Compression Goal has been Achieved

**[0130]** Finally, we compare the original number of bits (**384**) to the current number of bits ( $545=507+38$ ) that are needed for this compression pass. We find that it takes 141% or 1.41 times as many bits to store the compressed data as it took to store the original data. Compression is still not achieved and the amount of data in this technique is growing larger rather than smaller in comparison to the previous pass requiring 132%.

#### Pass 3

**[0131]** In this pass, we replace the most highly occurring tuple from the previous pass with a new symbol, and then we determine whether we have achieved our compression goal.

#### Identifying all Possible Tuples

**[0132]** We expand the tuple array **35"**. FIG. **28** by adding the symbol that was created in the previous pass. We add the symbol "3" as a first symbol and last symbol, and enter the tuples in the new cells.

#### Determining the Highly Occurring Tuple

**[0133]** The tuple array identifies the tuples that we look for and tally in our revised alphabet. In table **40"**, FIG. **29**, the Total Symbol Count is 232, and the tuple that occurs most frequently is the character 1 followed by character 3 (**1>3**). It occurs 48 times, which ties with the tuple of character 3 followed by character 0. We determine that the tuple **1>3** is the most complex tuple because it has a hypotenuse length  $25'$  of  $3.16 (\text{SQRT}(1^2+3^2))$ , and tuple **3>0** has a hypotenuse of  $3 (\text{SQRT}(0^2+3^2))$ .

#### Creating a Symbol for the Highly Occurring Tuple

**[0134]** We define a new symbol 4 to represent the most highly occurring tuple **1>3**, and add it to the dictionary **26'''**, FIG. **30**, for the alphabet that was developed in the previous passes.

#### Replacing the Tuple with the New Symbol

**[0135]** In the data stream, we replace every instance of the most highly occurring tuple from the earlier data stream with

the new single symbol. We replace the 48 instances of the 1>3 tuple with the symbol 4 and new data stream **30-4** is obtained, FIG. 31.

#### Encoding the Alphabet

**[0136]** We count the number of symbols in the data stream, and use the count to build a Huffman tree and code for the current alphabet as seen in table **41**", FIG. 32. There is no Huffman code assigned to the symbol 1 because there are no instances of this symbol in the compressed data in this pass. (This can be seen in the data stream **30-4**, FIG. 31.) The total symbol count has been reduced from 232 to 184 (e.g.,  $88+0+40+8+48$ , but not including the EOF marker).

#### Calculating the Compressed File Size

**[0137]** We need to evaluate whether our substitution reduces the amount of space that it takes to store the data. As seen in table **52**", FIG. 33, the total bits are equal to 340.

**[0138]** In table **57**", FIG. 34, the compression overhead in bits is 42.

#### Determining Whether the Compression Goal has been Achieved

**[0139]** Finally, we compare the original number of bits (384) to the current number of bits (382) that are needed for this compression pass. We find that it takes 0.99 times as many bits to store the compressed data as it took to store the original data. Compression is achieved.

#### Pass 4

**[0140]** In this pass, we replace the most highly occurring tuple from the previous pass with a new symbol, and then we determine whether we have achieved our compression goal.

#### Identifying all Possible Tuples

**[0141]** We expand the tuple array **35**", FIG. 36, by adding the symbol that was created in the previous pass. We add the symbol 4 as a first symbol and last symbol, and enter the tuples in the new cells.

#### Determining the Highly Occurring Tuple

**[0142]** The tuple array identifies the tuples that we look for and tally in our revised alphabet. In table **40**", FIG. 37, the Total Symbol Count=184 and the tuple that occurs most frequently is the character 4 followed by character 0 (4>0). It occurs 48 times.

#### Creating a Symbol for the Highly Occurring Tuple

**[0143]** We define a new symbol 5 to represent the 4>0 tuple, and add it to the dictionary **26-4**, FIG. 38, for the alphabet that was developed in the previous passes.

#### Replacing the Tuple with the New Symbol

**[0144]** In the data stream, we replace every instance of the most highly occurring tuple with the new single symbol. We replace the 48 instances of the 40 tuple in data stream **30-4**, FIG. 31, with the symbol 5 as seen in data stream **30-5**, FIG. 39.

#### Encoding the Alphabet

**[0145]** As demonstrated above, we count the number of symbols in the data stream, and use the count to build a Huffman tree and code for the current alphabet. There is no

Huffman code assigned to the symbol 1 and the symbol 4 because there are no instances of these symbols in the compressed data in this pass. The total symbol count has been reduced from 184 to 136 (e.g.,  $40+0+40+8+0+48$ , but not including the EOF marker) as seen in table **41-4**, FIG. 40.

#### Calculating the Compressed File Size

**[0146]** We need to evaluate whether our substitution reduces the amount of space that it takes to store the data. As seen in table **52**", FIG. 41, the total number of bits is 283.

**[0147]** As seen in table **57**", FIG. 42, the compression overhead in bits is 48.

#### Determining Whether the Compression Goal has been Achieved

**[0148]** Finally, we compare the original number of bits (384) to the current number of bits (331) that are needed for this compression pass as seen in table **58**", FIG. 43. In turn, we find that it takes 0.86 times as many bits to store the compressed data as it took to store the original data.

#### Pass 5

**[0149]** In this pass, we replace the most highly occurring tuple from the previous pass with a new symbol, and then we determine whether we have achieved our compression goal.

#### Identifying all Possible Tuples

**[0150]** We expand the tuple array by adding the symbol that was created in the previous pass. We add the symbol 5 as a first symbol and last symbol, and enter the tuples in the new cells as seen in table **35-4**, FIG. 44.

#### Determining the Highly Occurring Tuple

**[0151]** The tuple array identifies the tuples that we look for and tally in our revised alphabet as seen in table **40-4**, FIG. 45. (Total Symbol Count=136) The tuple that occurs most frequently is the symbol 2 followed by symbol 5 (2>5), which has a hypotenuse of 5.4. It occurs 39 times. This tuple ties with the tuple 0>2 (hypotenuse is 2) and 5>0 (hypotenuse is 5). The tuple 2>5 is the most complex based on the hypotenuse length 25" described above.

#### Creating a Symbol for the Highly Occurring Tuple

**[0152]** We define a new symbol 6 to represent the most highly occurring tuple 2>5, and add it to the dictionary for the alphabet that was developed in the previous passes as seen in table **26-5**, FIG. 46.

#### Replacing the Tuple with the New Symbol

**[0153]** In the data stream, we replace every instance of the most highly occurring tuple with the new single symbol. We replace the 39 instances of the 2>5 tuple in data stream **30-5**, FIG. 39, with the symbol 6 as seen in data stream **30-6**, FIG. 47.

#### Encoding the Alphabet

**[0154]** As demonstrated above, we count the number of symbols in the data stream, and use the count to build a Huffman tree and code for the current alphabet as seen in table **41-5**, FIG. 48. There is no Huffman code assigned to the symbol 1 and the symbol 4 because there are no instances of these symbols in the compressed data in this pass. The total

symbol count has been reduced from 136 to 97 (e.g.,  $40+1+8+9+39$ , but not including the EOF marker) as seen in table 52-4, FIG. 49.

#### Calculating the Compressed File Size

[0155] We need to evaluate whether our substitution reduces the amount of space that it takes to store the data. As seen in table 52-4, FIG. 49, the total number of bits is 187.

[0156] As seen in table 57-4, FIG. 50, the compression overhead in bits is 59.

Determining Whether the Compression Goal has been Achieved

[0157] Finally, we compare the original number of bits (384) to the current number of bits (246, or  $187+59$ ) that are needed for this compression pass as seen in table 58-4, FIG. 51. We find that it takes 0.64 times as many bits to store the compressed data as it took to store the original data.

#### Pass 6

[0158] In this pass, we replace the most highly occurring tuple from the previous pass with a new symbol, and then we determine whether we have achieved our compression goal.

#### Identifying all Possible Tuples

[0159] We expand the tuple array 35-5 by adding the symbol that was created in the previous pass as seen in FIG. 52. We add the symbol 6 as a first symbol and last symbol, and enter the tuples in the new cells.

#### Determining the Highly Occurring Tuple

[0160] The tuple array identifies the tuples that we look for and tally in our revised alphabet. (Total Symbol Count=97) The tuple that occurs most frequently is the symbol 0 followed by symbol 6 (0>6). It occurs 39 times as seen in table 40-5, FIG. 53.

#### Creating a Symbol for the Highly Occurring Tuple

[0161] We define a new symbol 7 to represent the 0>6 tuple, and add it to the dictionary for the alphabet that was developed in the previous passes as seen in table 26-6, FIG. 54.

#### Replacing the Topic with the New Symbol

[0162] In the data stream, we replace every instance of the most highly occurring tuple with the new single symbol. We replace the 39 instances of the 0>6 tuple in data stream 30-6, FIG. 47, with the symbol 7 as seen in data stream 30-7, FIG. 55.

#### Encoding the Alphabet

[0163] As demonstrated above, we count the number of symbols in the data stream, and use the count to build a Huffman tree and code for the current alphabet as seen in table 41-6, FIG. 56. There is no Huffman code assigned to the symbol 1, symbol 4 and symbol 6 because there are no instances of these symbols in the compressed data in this pass. The total symbol count has been reduced from 97 to 58 (e.g.,  $1+0+1+8+0+9+0+39$ , but not including the EOF marker).

[0164] Because all the symbols 1, 4, and 6 have been removed from the data stream, there is no reason to express them in the encoding scheme of the Huffman tree 50', FIG. 57. However, the extinct symbols will be needed in the decode table. A complex symbol may decode to two less complex symbols. For example, a symbol 7 decodes to 0>6.

[0165] We need to evaluate whether our substitution reduces the amount of space that it takes to store the data. As seen in table 52-5, FIG. 58, the total number of bits is 95.

[0166] As seen in table 57-5, FIG. 59, the compression overhead in bits is 71.

Determining Whether the Compression Goal has been Achieved

[0167] Finally, we compare the original number of bits (384) to the current number of bits (166, or  $95+71$ ) that are needed for this compression pass as seen in table 58-5, FIG. 60. We find that it takes 0.43 times as many bits to store the compressed data as it took to store the original data.

#### Subsequent Passes

[0168] Skilled artisans will also notice that overhead has been growing in size while the total number of bits is still decreasing. We repeat the procedure to determine if this is the optimum compressed file size. We compare the compression size for each subsequent pass to the first occurring lowest compressed file size. The chart 60, FIG. 61, demonstrates how the compressed file size grows, decreases, and then begins to grow as the encoding information and dictionary sizes grow. We can continue the compression of the foregoing techniques until the text file compresses to a single symbol after 27 passes.

#### Interesting Symbol Statistics

[0169] With reference to table 61, FIG. 62, interesting statistics about the symbols for this compression are observable. For instance, the top 8 symbols represent 384 bits (e.g.,  $312+45+24+2+1$ ) and 99.9% (e.g.,  $81.2+11.7+6.2+0.5+0.3\%$ ) of the file.

#### Storing the Compressed File

[0170] The information needed to decompress a file is usually written at the front of a compressed file, as well as to a separate dictionary only file. The compressed file contains information about the file, a coded representation of the Huffman tree that was used to compress the data, the dictionary of symbols that was created during the compression process, and the compressed data. The goal is to store the information and data in as few bits as possible.

[0171] This section describes a method and procedure for storing information in the compressed file.

#### File Type

[0172] The first four bits in the file are reserved for the version number of the file format, called the file type. This field allows flexibility for future versions of the software that might be used to write the encoded data to the storage media. The file type indicates which version of the software was used when we saved the file in order to allow the file to be decompressed later.

[0173] Four bits allows for up to 16 versions of the software. That is, binary numbers from 0000 to 1111 represent version numbers from 0 to 15. Currently, this field contains binary 0000.

#### Maximum Symbol Width

[0174] The second four bits in the file are reserved for the maximum symbol width. This is the number of bits that it takes to store in binary form the largest symbol value. The

actual value stored is four less than the number of bits required to store the largest symbol value in the compressed data. When we read the value, we add four to the stored number to get the actual maximum symbol width. This technique allows symbol values up to 20 bits. In practical terms, the value  $2^{20}$  (2 raised to the 20<sup>th</sup> power) means that about 1 million symbols can be used for encoding.

[0175] For example, if symbols 0-2000 might appear in the compressed file, the largest symbol ID (2000) would fit in a field containing 11 bits. Hence, a decimal 7 (binary 0111) would be stored in this field.

[0176] In the compression example, the maximum symbol width is the end-of-file symbol 8, which takes four bits in binary (1000). We subtract four, and store a value of 0000. When we decompress the data, we add four to zero to find the maximum symbol width of four bits. The symbol width is used to read the Huffman tree that immediately follows in the coded data stream.

#### Coded Huffman Tree

[0177] We must store the path information for each symbol that appears in the Huffman tree and its value. To do this, we convert the symbol's digital value to binary. Each symbol will be stored in the same number of bits, as determined by the symbol with the largest digital value and stored as the just read "symbol width".

[0178] In the example, the largest symbol in the dictionary in the Huffman encoded tree is the end-of-file symbol 8. The binary form of 8 is 1000, which takes 4 bits. We will store each of the symbol values in 4 bits.

[0179] To store a path, we will walk the Huffman tree in a method known as a pre-fix order recursive parse, where we visit each node of the tree in a known order. For each node in the tree one bit is stored. The value of the bit indicates if the node has children (1) or if it is a leaf with no children (0). If it is a leaf, we also store the symbol value. We start at the root and follow the left branch down first. We visit each node only once. When we return to the root, we follow the right branch down, and repeat the process for the right branch.

[0180] In the following example, the Huffman encoded tree is redrawn as 50-2 to illustrate the prefix-order parse, where nodes with children are labeled as 1, and leaf nodes are labeled as 0 as seen in FIG. 63.

[0181] The discovered paths and symbols are stored in the binary form in the order in which they are discovered in this method of parsing. Write the following bit string to the file, where the bits displayed in bold/underline represent the path, and the value of the 0 node are displayed without bold/underline. The spaces are added for readability; they are not written to media.

110 0101 110 0000 10 1000 0 0010 0 0011 0 0111

#### Encode Array for the Dictionary

[0182] The dictionary information is stored as sequential first/last definitions, starting with the two symbols that define the symbol 2. We can observe the following characteristics of the dictionary:

[0183] The symbols 0 and 1 are the atomic (non-divisible) symbols common to every compressed file, so they do not need to be written to media.

[0184] Because we know the symbols in the dictionary are sequential beginning with 2, we store only the symbol definition and not the symbol itself.

[0185] A symbol is defined by the tuple it replaces. The left and right symbols in the tuple are naturally symbols that precede the symbol they define in the dictionary.

[0186] We can store the left/right symbols of the tuple in binary form.

[0187] We can predict the maximum number of bits that it takes to store numbers in binary form. The number of bits used to store binary numbers increases by one bit with each additional power of two as seen, for example, in table 62, FIG. 64:

[0188] Because the symbol represents a tuple made up of lower-level symbols, we will increase the bit width at the next higher symbol value; that is, at 3, 5, 9, and 17, instead of at 2, 4, 8, and 16.

[0189] We use this information to minimize the amount of space needed to store the dictionary. We store the binary values for the tuple in the order of first and last, and use only the number of bits needed for the values.

[0190] Three dictionary instances have special meanings. The 0 and 1 symbols represent the atomic symbols of data binary 0 binary 1, respectively. The last structure in the array represents the end-of-file (EOF) symbol, which does not have any component pieces. The EOF symbol is always assigned a value that is one number higher than the last symbol found in the data stream.

[0191] Continuing our compression example, the table 63, FIG. 65, shows how the dictionary is stored.

[0192] Write the following bit string to the file. The spaces are added for readability; they are not written to media.

10 1000 0111 100000 010101 000110

#### Encoded Data

[0193] To store the encoded data, we replace the symbol with its matching Huffman code and write the bits to the media. At the end of the encoded bit string, we write the EOF symbol. In our example, the final compressed symbol string is seen again as 30-7, FIG. 66, including the EOF.

[0194] The Huffman code for the optimal compression is shown in table 67, FIG. 67.

[0195] As we step through the data stream, we replace the symbol with the Huffman coded bits as seen at string 68, FIG. 68. For example, we replace symbol 0 with the bits 0100 from table 67, replace symbol 5 with 00 from table 67, replace instances of symbol 7 with 1, and so on. We write the following string to the media, and write the end of file code at the end. The bits are separated by spaces for readability; the spaces are not written to media.

[0196] The compressed bit string for the data, without spaces is: 010000111111111111111111111111101010110110011111110110010100011000110001100011000101101010

#### Overview of the Stored File

[0197] As summarized in the diagram 69, FIG. 69, the information stored in the compressed file is the file type, symbol width, Huffman tree, dictionary, encoded data, and EOF symbol. After the EOF symbol, a variable amount of pad bits are added to align the data with the final byte in storage.

[0198] In the example, the bits 70 of FIG. 70 are written to media. Spaces are shown between the major fields for read-

ability; the spaces are not written to media. The “x” represents the pad bits. In FIG. 69, the bits 70 are seen filled into diagram 69b corresponding to the compressed file format.

#### Decompressing the Compressed File

**[0199]** The process of decompression unpacks the data from the beginning of the file 69, FIG. 69, to the end of the stream.

#### File Type

**[0200]** Read the first four bits of the file to determine the file format version.

#### Maximum Symbol Width

**[0201]** Read the next four bits in the file, and then add four to the value to determine the maximum symbol width. This value is needed to read the Huffman tree information.

#### Huffman Tree

**[0202]** Reconstruct the Huffman tree. Each 1 bit represents a node with two children. Each 0 bit represents a leaf node, and it is immediately followed by the symbol value. Read the number of bits for the symbol using the maximum symbol width.

**[0203]** In the example, the stored string for Huffman is:

**[0204]** 1100101110000010100000100001100111

**[0205]** With reference to FIG. 71, diagram 71 illustrates how to unpack and construct the Huffman tree using the pre-fix order method.

#### Dictionary

**[0206]** To reconstruct the dictionary from file 69, read the values for the pairs of tuples and populate the table. The values of 0 and 1 are known, so they are automatically included. The bits are read in groups based on the number of bits per symbol at that level as seen in table 72, FIG. 72.

**[0207]** In our example, the following bits were stored in the file: 1010000111101000010101000110

**[0208]** We read the numbers in pairs, according to the bits per symbol, where the pairs represent the numbers that define symbols in the dictionary:

Bits	Symbol
1 0	2
10 00	3
01 11	4
100 000	5
010 101	6
000 110	7

**[0209]** We convert each binary number to a decimal number:

Decimal Value	Symbol
1 0	2
2 0	3
1 3	4
4 0	5

-continued

Decimal Value	Symbol
2 5	6
0 6	7

**[0210]** We identify the decimal values as the tuple definitions for the symbols:

Symbol	Tuple
2	1 > 0
3	2 > 0
4	1 > 3
5	4 > 0
6	2 > 5
7	0 > 6

**[0211]** We populate the dictionary with these definitions as seen in table 73, FIG. 73.

#### Construct the Decode Tree

**[0212]** We use the tuples that are defined in the re-constructed dictionary to build the Huffman decode tree. Let's decode the example dictionary to demonstrate the process. The diagram 74 in FIG. 74 shows how we build the decode tree to determine the original bits represented by each of the symbols in the dictionary. The step-by-step reconstruction of the original bits is as follows:

**[0213]** Start with symbols 0 and 1. These are the atomic elements, so there is no related tuple. The symbol 0 is a left branch from the root. The symbol 1 is a right branch. (Left and right are relative to the node as you are facing the diagram—that is, on your left and on your right.) The atomic elements are each represented by a single bit, so the binary path and the original path are the same. Record the original bits 0 and 1 in the decode table.

**[0214]** Symbol 2 is defined as the tuple 1>0 (symbol 1 followed by symbol 0). In the decode tree, go to the node for symbol 1, then add a path that represents symbol 0. That is, add a left branch at node 1. The terminating node is the symbol 2. Traverse the path from the root to the leaf to read the branch paths of left (L) and right (R). Replace each left branch with a 0 and each right path with a 1 to view the binary form of the path as LR, or binary 10.

**[0215]** Symbol 3 is defined as the tuple 2>0. In the decode tree, go to the node for symbol 2, then add a path that represents symbol 0. That is, add a left branch at node 2. The terminating node is the symbol 3. Traverse the path from the root to the leaf to read the branch path of RLL. Replace each left branch with a 0 and each right path with a 1 to view the binary form of the path as 100.

**[0216]** Symbol 4 is defined as the tuple 1>3. In the decode tree, go to the node for symbol 1, then add a path that represents symbol 3. From the root to the node for symbol 3, the path is RLL. At symbol 1, add the RLL path. The terminating node is symbol 4. Traverse the path from the root to the leaf to read the path of RRLL, which translates to the binary format of 1100.

**[0217]** Symbol 5 is defined as the tuple 4>0. In the decode tree, go to the node for symbol 4, then add a path that represents symbol 0. At symbol 4, add the L path. The terminating

node is symbol 5. Traverse the path from the root to the leaf to read the path of RRLLL, which translates to the binary format of 11000.

[0218] Symbol 6 is defined as the tuple 2>5. In the decode tree, go to the node for symbol 2, then add a path that represents symbol 5. From the root to the node for symbol 5, the path is RRLLL. The terminating node is symbol 6. Traverse the path from the root to the leaf to read the path of RLRLLL, which translates to the binary format of 1011000.

[0219] Symbol 7 is defined as the tuple 0>6. In the decode tree, go to the node for symbol 0, then add a path that represents symbol 6. From the root to the node for symbol 6, the path is RLRLLL. The terminating node is symbol 7. Traverse the path from the root to the leaf to read the path of LRLRLLL, which translates to the binary format of 0101000.

#### Decompress the Data

[0220] To decompress the data, we need the reconstructed Huffman tree and the decode table that maps the symbols to their original bits as seen at 75, FIG. 75. We read the bits in the data file one bit at a time, following the branching path in the Huffman tree from the root to a node that represents a symbol. The compressed file data bits are: 0100001111111111111111111101100111011001111111011001011000110001100011000110101010

[0221] For example, the first four bits of encoded data 0100 takes us to symbol 0 in the Huffman tree, as illustrated in the diagram 76, FIG. 76. We look up 0 in the decode tree and table to find the original bits. In this case, the original bits are also 0. We replace 0100 with the single bit 0.

[0222] In the diagram 77 in FIG. 77, we follow the next two bits 00 to find symbol 5 in the Huffman tree. We look up 5 in the decode tree and table to find that symbol 5 represents original bits of 11000. We replace 00 with 11000.

[0223] In the diagram 78, FIG. 78, we follow the next bit 1 to find symbol 7 in the Huffman tree. We look up 7 in the decode tree and table to find that symbol 7 represents the original bits 01011000. We replace the single bit 1 with 01011000. We repeat this for each 1 in the series of 1s that follow.

[0224] The next symbol we discover is with bits 011. We follow these bits in the Huffman tree in diagram 79, FIG. 79. We look up symbol 3 in the decode tree and table to find that it represents original bits 100, so we replace 011 with bits 100.

[0225] We continue the decoding and replacement process to discover the symbol 2 near the end of the stream with bits 01011, as illustrated in diagram 80, FIG. 80. We look up symbol 2 in the decode tree and table to find that it represents original bits 10, so we replace 01011 with bits 10.

[0226] The final unique sequence of bits that we discover is the end-of-file sequence of 01010, as illustrated in diagram 81, FIG. 81. The EOF tells us that we are done unpacking.

[0227] Altogether, the unpacking of compressed bits recovers the original bits of the original data stream in the order of diagram 82 spread across two FIGS. 82a and 82b.

[0228] With reference to FIG. 83, a representative computing system environment 100 includes a computing device 120. Representatively, the device is a general or special purpose computer, a phone, a PDA, a server, a laptop, etc., having a hardware platform 128. The hardware platform includes physical I/O and platform devices, memory (M), processor (P), such as a CPU(s), USB or other interfaces (X), drivers (D), etc. In turn, the hardware platform hosts one or more

virtual machines in the form of domains 130-1 (domain 0, or management domain), 130-2 (domain U1), . . . 130-n (domain Un), each having its own guest operating system (O.S.) (e.g., Linux, Windows, Netware, Unix, etc.), applications 140-1, 140-2, . . . 140-n, file systems, etc. The workloads of each virtual machine also consume data stored on one or more disks 121.

[0229] An intervening Xen or other hypervisor layer 150, also known as a "virtual machine monitor" or virtualization manager, serves as a virtual interface to the hardware and virtualizes the hardware. It is also the lowest and most privileged layer and performs scheduling control between the virtual machines as they task the resources of the hardware platform, e.g., memory, processor, storage, network (N) (by way of network interface cards, for example), etc. The hypervisor also manages conflicts, among other things, caused by operating system access to privileged machine instructions. The hypervisor can also be type 1 (native) or type 2 (hosted). According to various partitions, the operating systems, applications, application data, boot data, or other data, executable instructions, etc., of the machines are virtually stored on the resources of the hardware platform. Alternatively, the computing system environment is not a virtual environment at all, but a more traditional environment lacking a hypervisor, and partitioned virtual domains. Also, the environment could include dedicated services or those hosted on other devices.

[0230] In any embodiment, the representative computing device 120 is arranged to communicate 180 with one or more other computing devices or networks. In this regard, the devices may use wired, wireless or combined connections to other devices/networks and may be direct or indirect connections. If direct, they typify connections within physical or network proximity (e.g., intranet). If indirect, they typify connections such as those found with the internet, satellites, radio transmissions, or the like. The connections may also be local area networks (LAN), wide area networks (WAN), metro area networks (MAN), etc., that are presented by way of example and not limitation. The topology is also any of a variety, such as ring, star, bridged, cascaded, meshed, or other known or hereinafter invented arrangement.

[0231] In still other embodiments, skilled artisans will appreciate that enterprises can implement some or all of the foregoing with humans, such as system administrators, computing devices, executable code, or combinations thereof. In turn, methods and apparatus of the invention further contemplate computer executable instructions, e.g., code or software, as part of computer program products on readable media, e.g., disks for insertion in a drive of a computing device 120, or available as downloads or direct use from an upstream computing device. When described in the context of such computer program products, it is denoted that items thereof, such as modules, routines, programs, objects, components, data structures, etc., perform particular tasks or implement particular abstract data types within various structures of the computing system which cause a certain function or group of function, and such are well known in the art.

[0232] While the foregoing produces a well-compressed output file, e.g., FIG. 69, skilled artisans should appreciate that the algorithm requires relatively considerable processing time to determine a Huffman tree, e.g., element 50, and a dictionary, e.g., element 26, of optimal symbols for use in encoding and compressing an original file. Also, the time spent to determine the key information of the file is significantly longer than the time spent to encode and compress the



file with the key. The following embodiment, therefore, describes a technique to use a file's compression byproducts to compress other data files that contain substantially similar patterns. The effectiveness of the resultant compression depends on how similar a related file's patterns are to the original file's patterns. As will be seen, using previously created, but related key, decreases the processing time to a small fraction of the time needed for the full process above, but at the expense of a slightly less effective compression. The process can be said to achieve a "fast approximation" to optimal compression for the related files.

[0233] The definitions from FIG. 1 still apply.

[0234] Broadly, the "fast approximation" hereafter 1) greatly reduces the processing time needed to compress a file using the techniques above, and 2) creates and uses a decode tree to identify the most complex possible pattern from an input bit stream that matches previously defined patterns. Similar to earlier embodiments, this encoding method requires repetitive computation that can be automated by computer software. The following discusses the logical processes involved.

#### Compression Procedure Using a Fast Approximation to Optimal Compression

[0235] Instead of using the iterative process of discovery of the optimal set of symbols, above, the following uses the symbols that were previously created for another file that contains patterns significantly similar to those of the file under consideration. In a high-level flow, the process involves the following tasks:

[0236] 1. Select a file that was previously compressed using the procedure(s) in FIGS. 2-82*b*. The file should contain data patterns that are significantly similar to the current file under consideration for compression.

[0237] 2. From the previously compressed file, read its key information and unpack its Huffman tree and symbol dictionary by using the procedure described above, e.g., FIGS. 63-82*b*.

[0238] 3. Create a decode tree for the current file by using the symbol dictionary from the original file.

[0239] 4. Identify and count the number of occurrences of patterns in the current file that match the previously defined patterns.

[0240] 5. Create a Huffman encoding tree for the symbols that occur in the current file plus an end-of-file (EOF) symbol.

[0241] 6. Store the information using the Huffman tree for the current file plus the file type, symbol width, and dictionary from the original file.

Each of the tasks is described in more detail below. An example is provided thereafter.

#### Selecting a Previously Compressed File

[0242] The objective of the fast approximation method is to take advantage of the key information in an optimally compressed file that was created by using the techniques above. In its uncompressed form of original data, the compressed file should contain data patterns that are significantly similar to the patterns in the current file under consideration for compression. The effectiveness of the resultant compression depends on how similar a related file's patterns are to the original file's patterns. The way a skilled artisan recognizes a similar file is that similar bit patterns are found in the origi-

nally compressed and new file yet to be compressed. It can be theorized a priori that files are likely similar if they have similar formatting (e.g., text, audio, image, powerpoint, spreadsheet, etc), topic content, tools used to create the files, file type, etc. Conclusive evidence of similar bit patterns is that similar compression ratios will occur on both files (i.e. original file compresses to 35% of original size, while target file also compresses to about 35% of original size). It should be noted that similar file sizes are not a requisite for similar patterns being present in both files.

[0243] With reference to FIG. 84, the key information 200 of a file includes the file type, symbol width, Huffman tree, and dictionary from an earlier file, e.g., file 69, FIG. 69.

#### Reading and Unpacking the Key Information

[0244] From the key information 200, read and unpack the File Type, Maximum Symbol Width, Huffman Tree, and Dictionary fields.

#### Creating a Decode Tree for the Current File

[0245] Create a pattern decode tree using the symbol dictionary retrieved from the key information. Each symbol represents a bit pattern from the original data stream. We determine what those bits are by building a decode tree, and then parsing the tree to read the bit patterns for each symbol.

[0246] We use the tuples that are defined in the re-constructed dictionary to build the decode tree. The pattern decode tree is formed as a tree that begins at the root and branches downward. A terminal node represents a symbol ID value. A transition node is a placeholder for a bit that leads to terminal nodes.

#### Identifying and Counting Pattern Occurrences

[0247] Read the bit stream of the current file one bit at a time. As the data stream is parsed from left to right, the paths in the decode tree are traversed to detect patterns in the data that match symbols in the original dictionary.

[0248] Starting from the root of the pattern decode tree, use the value of each input bit to determine the descent path thru the pattern decode tree. A "0" indicates a path down and to the left, while a "1" indicates a path down and to the right. Continue descending through the decode tree until there is no more descent path available. This can occur because a branch left is indicated with no left branch available, or a branch right is indicated with no right branch available.

[0249] When the end of the descent path is reached, one of the following occurs:

[0250] If the descent path ends in a terminal node, count the symbol ID found there.

[0251] If the descent path ends in a transition node, retrace the descent path toward the root, until a terminal node is encountered. This terminal node represents the most complex pattern that could be identified in the input bit stream. For each level of the tree ascended, replace the bit that the path represents back into the bit stream because those bits form the beginning of the next pattern to be discovered. Count the symbol ID found in the terminal node.

[0252] Return to the root of the decode tree and continue with the next bit in the data stream to find the next symbol.

[0253] Repeat this process until all of the bits in the stream have been matched to patterns in the decode tree. When done,





coded to Huffman codes. For example, the “0” bit at position 250 in the original bit stream coded to a symbol “0” as described in FIG. 88. By replacing the symbol0 with its Huffman code (1001) from table 290, FIG. 91, the Huffman encoded bits are seen, as: 1001 0 11 0 11 0 0 11 0 11 0 0 11 0 11 0 11 0 11 11 11 11 0 0 11 0 11 0 1011 1010 1000

[0284] Spaces are shown between the coded bits for readability; the spaces are not written to media. Also, the code for the EOF symbol (1000) is placed at the end of the encoded data and shown in underline.

[0285] With reference to FIG. 93, the foregoing information is stored in the compressed file 69' for the current file. As skilled artisans will notice, it includes both original or re-used information and new information, thereby resulting in a “fast approximation.” In detail, it includes the file type from the original key information (200), the symbol width from the original key information (200), the new Huffman coding recently created for the new file, the dictionary from the key information (200) of the original file, the data that is encoded by using the new Huffman tree, and the new EOF symbol. After the EOF symbol, a variable amount of pad bits are added to align the data with the final byte in storage.

[0286] In still another alternate embodiment, the following describes technology to identify a file by its contents. It is defined, in one sense, as providing a file’s “digital spectrum.” The spectrum, in turn, is used to define a file’s position in an N-dimensional universe. This universe provides a basis by which a file’s position determines similarity, adjacency, differentiation and grouping relative to other files. Ultimately, similar files can originate many new compression features, such as the “fast approximations” described above. The terminology defined in FIG. 1 remains valid as does the earlier-presented information for compression and/or fast approximations using similar files. It is supplemented with the definitions in FIG. 94. Also, the following considers an alternate use of the earlier described symbols to define a digital variance in a file. For simplicity in this embodiment, a data stream under consideration is sometimes referred to as a “file.”

[0287] The set of values that digitally identifies the file, referred to as the file’s digital spectrum, consists of several pieces of information found in two scalar values and two vectors.

The scalar values are:

[0288] The number of symbols in the symbol dictionary (the dictionary being previously determined above.)

[0289] The number of symbols also represents the number of dimensions in the N-dimensional universe, and thus, the number of coordinates in the vectors.

[0290] The length of the source file in bits.

[0291] This is the total number of bits in the symbolized data stream after replacing each symbol with the original bits that the symbol represents.

The Vectors are:

[0292] An ordered vector of frequency counts, where each count represents the number of times a particular symbol is detected in the symbolized data stream.

[0293]  $F_x = (F_{0x}, F_{1x}, F_{2x}, F_{3x}, \dots, F_{Nx})$ ,

where F represents the symbol frequency vector, 0 to N are the symbols in a file’s symbol dictionary, and x represents the source file of interest.

[0294] An ordered vector of bit lengths, where each bit length represents the number of bits that are represented by a particular symbol.

[0295]  $B_x = (B_{0x}, B_{1x}, B_{2x}, B_{3x}, \dots, B_{Nx})$ ,

[0296] where B represents the bit-length vector, 0 to N are the symbols in a file’s symbol dictionary, and x represents the source file of interest.

[0297] The symbol frequency vector can be thought of as a series of coordinates in an N-dimensional universe where N is the number of symbols defined in the alphabet of the dictionary, and the counts represent the distance from the origin along the related coordinate axis. The vector describes the file’s informational position in the N-dimension universe. The meaning of each dimension is defined by the meaning of its respective symbol.

[0298] The origin of N-dimensional space is an ordered vector with a value of 0 for each coordinate:

[0299]  $F_o = (0, 0, 0, 0, 0, 0, 0, \dots, 0)$ .

[0300] The magnitude of the frequency vector is calculated relative to the origin. An azimuth in each dimension can also be determined using ordinary trigonometry, which may be used at a later time. By using Pythagorean geometry, the distance from the origin to any point F, in the N-dimensional space can be calculated, i.e.:

$$D_{ox} = \text{square root}(((F_{0x} - F_{0o})^2) + ((F_{1x} - F_{1o})^2) + ((F_{2x} - F_{2o})^2) + ((F_{3x} - F_{3o})^2) + \dots + ((F_{Nx} - F_{No})^2))$$

[0301] Substituting the 0 at each coordinate for the values at the origin, the simplified equation is:

$$D_{ox} = \text{square root}((F_{0x})^2 + (F_{1x})^2 + (F_{2x})^2 + (F_{3x})^2 + \dots + (F_{Nx})^2)$$

[0302] As an example, imagine that a file has 10 possible symbols and the frequency vector for the file is:

$$F_x = (3, 5, 6, 1, 0, 7, 19, 3, 6, 22).$$

[0303] Since this vector also describes the file’s informational position in this 10-dimension universe, its distance from the origin can be calculated using the geometry outlined. Namely,:

$$D_{ox} = \text{square root}(((3-0)^2) + ((5-0)^2) + ((6-0)^2) + ((6-0)^2) + ((1-0)^2) + ((0-0)^2) + ((7-0)^2) + ((19-0)^2) + ((3-0)^2) + ((22-0)^2))$$

[0304]  $D_{ox} = 31.78$ .

Determining a Characteristic Digital Spectrum

[0305] To create a digital spectrum for a file under current consideration, we begin with the key information 200, FIG. 84, which resulted from an original file of interest. The digital spectrum determined for this original file is referred to as the characteristic digital spectrum. A digital spectrum for a related file of interest, on the other hand, is determined by its key information from another file. Its digital spectrum is referred to as a related digital spectrum.

[0306] The key information actually selected for the characteristic digital spectrum is considered to be a “well-suited key.” A “well-suited key” is a key best derived from original data that is substantially similar to the current data in a current file or source file to be examined. The key might even be the actual compression key for the source file under consideration. However, to eventually use the digital spectrum information for the purpose of file comparisons and grouping, it is necessary to use a key that is not optimal for any specific file, but that can be used to define the N-dimensional symbol

universe in which all the files of interest are positioned and compared. The more closely a key matches a majority of the files to be examined, the more meaningful it is during subsequent comparisons.

**[0307]** The well-suited key can be used to derive the digital spectrum information for the characteristic file that we use to define the N-dimensional universe in which we will analyze the digital spectra of other files. From above, the following information is known about the characteristic digital spectrum of the file:

[0308] The number of symbols (N) in the symbol dictionary

[0309] The length of the source file in bits

**[0310]** An ordered vector of symbol frequency counts

[0311]  $F_i = (F_{0i}, F_{1i}, F_{2i}, F_{3i}, \dots, F_{Ni})$ ,

[0312] where F represents the symbol frequency, 0 to N are the symbols in the characteristic file's symbol dictionary, and i represents the characteristic file of interest.

[0313] An ordered vector of bit lengths

[0314]  $B_i = (B_{0i}, B_{1i}, B_{2i}, B_{3i}, \dots, B_{Ni}),$

[0315] where B represents the bit-length vector, 0 to N are the symbols in the characteristic file's symbol dictionary, and i represents the characteristic file of interest.

### Determining a Related Digital Spectrum

**[0316]** Using the key information and digital spectrum of the characteristic file, execute the process described in the fast approximation embodiment for a current, related file of interest, but with the following changes:

**[0317]** 1. Create a symbol frequency vector that contains one coordinate position for the set of symbols described in the characteristic file's symbol dictionary.

**[0318]**  $F_i = (F_{0i}, F_{1i}, F_{2i}, F_{3i}, \dots, F_{Ni}),$

[0319] where F represents the symbol frequency, 0 to N are the symbols in the characteristic file's symbol dictionary, and j represents the related file of interest.

**[0320]** Initially, the count for each symbol is zero (0).

[0321] 2. Parse the data stream of the related file of interest for symbols. As the file is parsed, conduct the following:

**[0322]**    a. Tally the instance of each discovered symbol in its corresponding coordinate position in the symbol frequency vector. That is, increment the respective counter for a symbol each time it is detected in the source file.

**[0323]** b. Do not Huffman encode or write the detected symbol.

[0324] c. Continue parsing until the end of the file is reached.

**[0325]** 3. At the completion of the source file parsing, write a digital spectrum output file that contains the following:

**[0326]** a. The number of symbols (N) in the symbol dictionary

[0327] b. The length of the source file in bits

[0328] c. The symbol frequency vector developed in the previous steps.

[0329]  $F_i = (F_{Di}, F_{1i}, F_{2i}, F_{3i}, \dots, F_{Ni}),$

[0330] where  $F$  represents the frequency vector, 0 to  $N$  are the symbols in the characteristic file's symbol dictionary, and the  $j$  represents the file of interest.

**[0331]** d. The bit length vector

[0332]  $B_i = (B_{0i}, B_{1i}, B_{2i}, B_{3i}, \dots, B_{Ni}),$

[0333] where B represents the bit-length vector, 0 to N are the symbols in the characteristic file's symbol dictionary, and j represents the file of interest.

### Advantages of Digital Spectrum Analysis

**[0334]** The digital spectrum of a file can be used to catalog a file's position in an N-dimensional space. This position in space, or digital spectrum, can be used to compute "distances" between file positions, and hence similarity, e.g., the closer the distance, the closer the similarity. The notion of a digital spectrum may eventually lead to the notion of a self-cataloging capability of digital files, or other.

## Begin: Example Defining a File's Digital Spectrum

**[0335]** To demonstrate the foregoing embodiment, the digital spectrum will be determined for a small data file that contains the following simple ASCII characters:

$$\text{aaaaaaaaaaaaaaaaaaaaaaaaaaaaabaaaabaaaaaaaaabbbbbbb} \quad (\text{eqn. 100})$$

**[0336]** Each character is stored as a sequence of eight bits that correlates to the ASCII code assigned to the character. The bit values for each character are:

$$a=01100001 \quad (\text{eqn. 101})$$
$$b=01100010 \quad (\text{eqn. 102})$$

**[0337]** By substituting the bits of equations 101 and 102 for the “a” and “b” characters in equation 100, a data stream **30** results as seen in FIG. 9. (Again, the characters are separated in the Figure with spaces for readability, but the spaces are not considered, just the characters.)

[10338] After performing an optimal compression of the data by using the process defined above in early embodiments, the symbols remaining in the data stream 30-7 are seen in FIG. 55. Alternatively, they are shown here as:

$$\begin{array}{l}05777777777777777777 \\ 35777357777777357353535352\end{array}\quad (\text{eqn. } 103)$$

**[0339]** With reference to FIG. 95, table 300 identifies the symbol definitions from equation 103 and the bits they represent. The symbol definition 302 identifies the alphabet of symbols determined from the data during the compression process. The symbols 0 and 1 are atomic symbols and represent original bits 0 and 1, by definition. The subsequent symbols, i.e., 2-7, are defined by tuples, or ordered pairs of symbols, that are represented in the data, e.g., symbol 4 corresponds to a “1” followed by a “3,” or 1>3. In turn, each symbol represents a series or sequence of bits 304 in the data stream of equation 103 (the source file), e.g., symbol 4 corresponds to original bits 1100.

**[0340]** With reference to table **310**, FIG. **96**, the number of occurrences of each symbol is counted in the data stream (equation 103) and the number of bits represented by each symbol is counted. For example, the symbol “7” in equation 103 appears thirty nine (39) times. In that its original bits **304**, correspond to “01011000,” it has eight (8) original bits appearing in the data stream for every instance of a “symbol 7” appearing. For a grand total of numbers of bits, the symbol count **312** is multiplied by the bit length **314** to arrive at a bit count **316**. In this instance, thirty nine (39) is multiplied by eight (8) to achieve a bit count of three-hundred twelve (312) for the symbol 7. A grand total of the number of bit counts **316**

for every symbol **320** gives a length of the source file **325** in numbers of bits. In this instance, the source file length (in bits) is three-hundred eighty-four (384).

[0341] In turn, the scalar values to be used in the file's digital spectrum are:

[0342] Source File Length in bits=384

[0343] Number of Symbols=8 total (or symbols 0 through 7, column **320**, FIG. **96**)

The vectors to be used in the file's digital spectrum are:

[0344] Frequency spectrum,  $F_x$ , represented by the ordered vector of counts for each symbol, from column **312**, FIG. **96**:

[0345]  $F_x=(1, 0, 1, 8, 0, 9, 0, 39)$

[0346] Bit length spectrum,  $B_x$ , is represented by the ordered vector of counts for the original bits in the file that are represented by each symbol, from column **314**, FIG. **96**:

[0347]  $B_x=(1, 1, 2, 3, 4, 5, 7, 8)$

[0348] The digital spectrum information can be used to calculate various useful characteristics regarding the file from which it was derived, as well as its relationship to other spectra, and the files from which the other spectra were derived. As an example, the frequency spectrum  $F(x)$  shown above, may be thought to describe a file's informational position in an 8-dimension universe, where the meaning of each dimension is defined by the meaning of its respective symbols.

[0349] Since the origin of the 8-dimensional space is an ordered vector with a value of 0 at each symbol position, e.g.,  $F(0)=(0,0,0,0,0,0,0,0)$ , the informational position in 8-dimensional space can be defined as an azimuth and distance from the origin. The magnitude of the position vector is calculated using Pythagorean geometry.  $\text{Dist}(x,0)=\sqrt{((F(x,0)-F(0,0))^2+\dots+(F(x,7)-F(0,7))^2)}$ . Simplified, this magnitude becomes  $\text{Dist}(x,0)=\sqrt{(F(x,0)^2+F(x,2)^2+F(x,3)^2+\dots+F(x,7)^2)}$ . Using the values above in  $F_x$ , the magnitude of the  $\text{Dist}(x,0)=40.84$ , or  $D_{x0}=\text{square root } (((1)^2+((0)^2)+(1)^2+((8)^2)+((0)^2)+(9)^2)+((0)^2)+(39)^2)=\text{square root } (1+0+1+64+0+81+0+1521)=40.84$ . Azimuth of the vector can be computed using basic trigonometry. Comparison of computed positions between files is useful to determine similarity, or not, of two or more subject files.

[0350] Another way to use the digital spectrum is to consider the vectors as defining points of a line graph **350**, as presented in FIG. **97A**. As an example, the X axis, labeled as "Patterns," defines a position for each symbol (e.g., element **320**, FIG. **96**) represented in the Frequency Spectrum. The Y axis, labeled as "Frequency," defines the values of the frequencies (or derivatives thereof) found in the Frequency Spectrum. From FIG. **96**, for example, symbols "3" and "5" have counts or frequencies of eight (8) and nine (9), respectively, and would be plotted in the line graph at x-y coordinates of (3,8) and (5,9).

[0351] Determination of the similarity of two digital spectra can be representatively determined using standard least squares statistical curve fitting techniques. In FIGS. **97A** and **97B**, four digital spectra are presented for comparison, whereby:

[0352] File 1 has a frequency spectrum  $F1=(1,4,13,5,12,6,20,15,18,21)$ ;

[0353] File 2 has a frequency spectrum  $F2=(1,5,13,6,15,5,21,20,15,20)$ ;

[0354] File 3 has a frequency spectrum  $F3=(2,9,8,9,21,10,15,10,15,24)$ ; and

[0355] File 4 has a frequency spectrum  $F4=(3,10,7,9,22,12,15,12,16,25)$ .

[0356] As best seen in FIG. **98B**, skilled artisans will visually recognize that the graph represented by F1 is "closest" to the graph of F2. Similarly, the graphs represented by F3 and F4 are closer to each other than, for instance, than either of the graphs represented by spectra from F1 and F2. "Closeness" is seen in the figure by graph filler **351** and **352** in the area between the files. For mathematical comparison purposes, a notion of the area between a reference and target graph can be determined. In a representative embodiment, a suitable method, which minimizes small differences and accentuates larger differences, is a sum of the squares of the differences at each point. The measurement of a difference function between two graphs, File x and File y, is computed as follows:

$$D(x,y)=((Fx1-Fy1)^2+(Fx2-Fy2)^2+(Fx3-Fy3)^2+\dots+(FxN-FyN)^2).$$

[0357] Hence, for the representation of the difference function between the above files F1 and F2, the computation is:

$$D(F1,F2)=((1-1)^2+(4-5)^2+(13-13)^2+(5-6)^2+(12-15)^2+(6-6)^2+(20-21)^2+(15-20)^2+(18-15)^2+(21-20)^2=48.$$

[0358] A representation of the difference function between files F1 and F3 is:

$$D(F1,F3)=((1-2)^2+(4-9)^2+(13-8)^2+(5-9)^2+(12-21)^2+(6-10)^2+(20-15)^2+(15-10)^2+(18-15)^2+(21-24)^2=232.$$

[0359] A matrix of the value of difference functions between each possible spectra graph may be computed to determine a measure of closeness between each possible spectra pair. The difference function values matrix for the above set of spectra F1, F2, F3 and F4 using the example comparison technique looks like this:

Spectra ID	F1	F2	F3	F4
F1	0	48	232	282
F2	48	0	264	298
F3	232	264	0	14
F4	282	298	14	0

[0360] Examination of the difference values between spectra lead to many useful conclusions. For example, the two most closely similar spectra are those belonging to files F3 and F4 (difference function value 14). The most dissimilar spectra, on the other hand, are those corresponding to files F2 and F4 (difference value 298). Relative to spectrum ID F1, the most similar spectrum is that belonging to file F2. Relative to spectrum ID F2, the most similar spectrum is that belonging to file F1. Relative to spectrum ID F3, the most similar spectrum is that belonging to file F4. Relative to spectrum ID F4, the most similar spectrum is that belonging to file F3. It can be also observed that there seem to be two groups of files, with two files in each group; files F1 and F2; the other group is files F3 and F4. The "closeness" of the latter group (e.g., difference value=14) is a much "tighter" grouping than that for the former group (e.g., difference value of 48). Skilled artisans will readily recognize the usefulness of these characteristics as a springboard for ascertaining still other properties and manipulations of files.

[0361] The foregoing has been described in terms of specific embodiments, but one of ordinary skill in the art will

recognize that additional embodiments are possible without departing from its teachings. This detailed description, therefore, and particularly the specific details of the exemplary embodiments disclosed, is given primarily for clarity of understanding, and no unnecessary limitations are to be implied, for modifications will become evident to those skilled in the art upon reading this disclosure and may be made without departing from the spirit or scope of the invention. Relatively apparent modifications, of course, include combining the various features of one or more figures with the features of one or more of the other figures.

1. In a computing system environment, a method of determining a digital spectrum of a file stored on a computing device, the file having a plurality of symbols representing an underlying data stream of original bits of data, comprising determining a number of occurrences of each said symbol in the file.

2. The method of claim 1, further including determining how many different symbols (N) are in the plurality of symbols thereby defining an N-dimensional space.

3. The method of claim 1, further including determining a magnitude of the file based on the determined number of occurrences

4. The method of claim 3, further including determining similarity of the file to another file by computing a distance function between the magnitudes of the file and the another file.

5. The method of claim 1, further including determining a number of the original bits of data represented by an entirety of the determined number of occurrences of said each symbol in the file.

6. The method of claim 1, further including determining each of the original bits of data for every symbol of the plurality of symbols.

7. In a computing system environment, a method of determining a digital spectrum of a file stored on a computing device, the file having a plurality of symbols, comprising:  
determining original bits of data for every said symbol; and  
determining a number of occurrences of said every symbol in the file.

8. The method of claim 7, further including determining how many different symbols (N) are in the plurality of symbols thereby defining an N-dimensional space, the deter-

mined number of occurrences being used to create an ordered vector in the N-dimensional space.

9. The method of claim 8, further including determining a magnitude of the file based on the ordered vector relative to an origin of the N-dimensional space.

10. The method of claim 9, further including determining a number of occurrences of every symbol in a second file.

11. The method of claim 10, further including creating a second ordered vector in the N-dimensional space for the second file.

12. The method of claim 11, further including determining a second magnitude of the second file relative to the origin of the N-Dimensional space.

13. The method of claim 13, further including comparing the second magnitude of the second file to the magnitude of the file to determine similarity of the files.

14. In a computing system environment, a method of determining similarity of two or more files stored on one or more computing devices, each said file having a plurality of symbols representing an underlying data stream of original bits of data, comprising:

determining a number of occurrences of every said symbol in said each file; and

comparing the number of occurrences between the files.

15. The method of claim 14, wherein said each file has a same total number of different symbols (N) thereby defining an N-dimensional space, the determined number of occurrences being used to create an ordered vector in the N-dimensional space for said each file.

16. The method of claim 15, further including determining a magnitude of said each file based on the ordered vectors relative to an origin of the N-dimensional space.

17. The method of claim 16, wherein the comparing further includes comparing the magnitudes of said each file.

18. The method of claim 14, further including determining for said each file a total number of the original bits of data.

19. The method of claim 18, further including determining for said each file a series of bits representing each of the original bits of data for said every symbol of the plurality of symbols.

20. A computer program product having executable instructions for loading on a computing device that undertake the method of claim 1.

\* \* \* \* \*